

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

DEEP BLUE

Rare type of diamond formed when boron hitchhikes to Earth's lower mantle on oceanic tectonic plates PAGES 35 & 84

STANDARD MODEL

HIGH-ENERGY FRONTIER

How machine learning is transforming particle physics

PAGE 41

METABOLISM

WEIGHT CONTROL

Succinate triggers calorie burning in brown fat

PAGES 38 & 102

EVOLUTION

WINDS OF CHANGE

Hurricanes favour lizards with strong grip on branches

PAGE 88

NATURE.COM/NATURE

2 August 2018

Vol. 560, No. 7716

THIS WEEK

EDITORIALS

WORLD VIEW Gather data to measure the impact of wildfires **p.7**

QUICK THINKING How quantum computers could battle for supremacy **p.8**

ANTS The genetic switch that makes minions a queen **p.9**



How lizards got their big feet

Creatures that survived wild weather in the Caribbean offer a rare and enlightening glimpse of natural selection in action.

When Charles Darwin first saw a Madagascar star orchid — sent to him by an enthusiast — he predicted the existence of a long-tongued pollinator that could reach for nectar inside the flowers' long tubes. The discovery of Morgan's sphinx moth, which had a tongue just long enough (and no longer), proved Darwin right — some two decades after his death.

It's one of the great demonstrations of evolution by natural selection. But what naturalists really want is to catch natural selection in the act, as it's doing the selecting. And opportunity knocked at the door of biologist Colin Donihue at Harvard University in Cambridge, Massachusetts, a year ago, just after he and his colleagues had returned from the Turks and Caicos Islands, where they had been studying anole lizards (*Anolis scriptus*).

On 8 September last year, Hurricane Irma struck the islands, battering them with sustained winds of up to 265 kilometres per hour. Two weeks later, Hurricane Maria arrived. Dozens of people in the region died. Reconstruction and rebuilding efforts continue.

Three weeks after the winds had died down, the researchers went back to assess the damage, and to see how (or even if) their lizards had survived. Their study, published this week in *Nature*, is the first to use an immediate before-and-after comparison to assess the impacts of hurricanes on evolutionary selection (C. M. Donihue *et al. Nature* http://doi.org/csgp; 2018). They saw clear trends of natural selection in action. In general, anoles found after the storms had bigger toepads, longer forelimbs and shorter hindlimbs than did lizards collected before the storm.

What have these traits to do with hurricanes? The lizards live in bushes and other low-growing vegetation. Toepads allow them purchase on the branches as they move, and it's a fair bet that limb proportions also play a part in helping a lizard to keep in contact with a branch, resisting moves by predators, other lizards — or, as it turned out, hurricanes — to knock them off.

The researchers took their idea further with a simple laboratory experiment in which they allowed lizards to get settled on a perch and then blew them off using a commercial leaf blower. (The lizards flew into comfy padding, and were not injured in the experiments.)

This experiment showed that when lizards are subjected to a stiff breeze, they hang on tightly with their forelimbs and let their hindlimbs hang loose. Longer hindlimbs, then, offer more purchase, explaining why lizards found after the storms tended to have shorter hindlimbs, but longer forelimbs.

These bigger toepads, shorter hindlimbs and longer forelimbs did not evolve as a direct response to the hurricanes. Natural selection interfered with the way in which these traits were spread across the population. Specifically, those lizards unable to hang on when the storms blew up — those with smaller toepads, longer hindlimbs and shorter forelimbs — were (presumably) blown away and perished. Those lizards better able to hang on tightly would have survived to weather another day. In technical terms, the mean values of the crucial traits measured before the storms had shifted.

These changes are phenotypic — merely observable characteristics. They say nothing about the genetic assimilation of such changes, which will presumably happen when the surviving lizards breed and new lizards are recruited to the population. Such changes are unlikely to be the last, given the extreme weather expected in the future. ■

Hot topic

Pinning extreme weather on climate change is now routine and reliable science.

Extreme weather is grabbing public attention. And a familiar question is being asked: is climate change to blame?

For years, the standard response was that climate change makes such events more likely, but it is hard to pin down the causes of a particular event. That is now changing.

As we highlight in a News Feature this week (see page 20), extreme-event attribution — the science of calculating how global warming has changed the likelihood and magnitude of extreme heat, cold, drought, rain or flooding — is ready to leave the lab. The research has advanced to the point at which public agencies can take over the task. That's down to progress in modelling, and more-accurate observations.

Rapid-attribution services are already being set up in Germany and elsewhere in Europe. Officials on other continents should follow suit.

Climate-change attribution statements are not yet ready for prime-time weather reports. But attribution is not just for the media or public curiosity. Policymakers, risk managers and courts will have a new decision-making tool on hand for issues related to climate change — to help plan infrastructure and, ultimately, assess liability.

Scientists should continue to improve the usefulness of attribution statements. There is no universal threshold beyond which climate change becomes dangerous, so research must establish resilience limits for different regions and sectors of the economy. To do this, and to help prepare a large city's health system or a region's agriculture for more turbulent weather, climate researchers must join forces with economists, social scientists and local specialists.

As attribution proceeds from science to service, the researchers who developed it can return to climate physics, such as exploring how clouds and climate interact, and how large-scale atmospheric circulation will change in a warming world. This is needed to continue checking the output of the models and to make attribution more precise. That can only help to highlight that our world is changing and we need to respond. ■



MATTHEW NEWTON/UNIV. TASMANIA

Wildfire science is at a loss for comprehensive data

An international monitoring initiative is crucial for understanding wildfires and reducing their damage, says **David Bowman**.

In 2013, a ferocious fire destroyed a small seaside village east of Hobart, Australia, the city where I live and work as a pyrogeographer. A newspaper photograph captured the terror of the firestorm: in it, a family shelters under a jetty in surreal, smoky orange twilight. Last week, spookily similar images from Greece, where people fled to the sea to escape an inferno, hit me with a jolt.

Fires have ignited elsewhere, including in Sweden and California, in recent weeks. But we can say little for certain about trends in wildfires worldwide. Data are too scant to say conclusively whether fires are becoming more destructive. If humans are to live sustainably on flammable landscapes, we need a global system for collecting data on fires to gain a coherent picture and assess strategies.

My colleagues and I analysed records from 1979 to 2013 (W. M. Jolly *et al. Nature Commun.* **6**, 7537; 2015). We found that fire seasons worldwide are lengthening, and that 'fire weather' — the combination of humidity, temperature, wind and other factors that help blazes to spread — is becoming more extreme. Yet the strong links between humans and flammable landscapes make fire a natural hazard like no other. We can both amplify and dampen the cycle by setting fires and fighting them. So a longer fire season does not necessarily mean more fires.

Satellite imaging has revolutionized our understanding of fire activity: it has provided global data on areas burnt and on variations from season to season regarding when fires occur and how large they grow. But satellite imagery is imperfect. Time courses of high-resolution images are separated by days, weeks or even months, if clouds cover the fires or patches of burnt ground. Geostationary satellites, which have higher orbits than other types, can provide real-time information, but only at coarse resolution.

In addition, satellite images became widely available only in the 1980s, so scientists know little about areas that burn infrequently. Instead, we rely on historical proxies, such as dating fire scars on tree trunks and examining charcoal layers in lake sediments.

Satellite analysis suggests that the total area burnt by wildfires over the past 18 years has declined (N. Andela *et al. Science* **356**, 1356–1362; 2017). This seems to be, in part, because much of the world's tropical savanna has been converted to ranches and cropland that fragments landscapes and makes them less flammable. For disasters, however, area burnt is usually less important than fire intensity, as determined by estimates of the energy that fires release. We've found that disastrous fires typically occur in regions with medium population densities, and at times of anomalously high heat, wind or dryness (D. M. J. S. Bowman *et al. Nature Ecol. Evol.* **1**, 0058; 2017).

Even a low-intensity fire can have a severe biological impact, especially in areas, such as rainforest, that are poorly adapted to handle it. And reliably assessing severity requires field work to collect data that might otherwise be missed. For example, fires under forest canopies are generally invisible to satellites.

To gain a comprehensive view, we need an initiative similar to the confederation of national meteorological networks that monitors daily weather conditions. This global network, established over the nineteenth and twentieth centuries, is an underappreciated scientific triumph. Its data are the backbone of weather forecasting and calibrating climate-change models. Imagine how unreliable weather projections would be if based only on satellite data, spotty field measurements and historical reconstructions — as is the case for landscape fires.

Comparative analysis of fire activity became possible in Europe only in 2004, with the advent of the European Union's fire database, which contains data from 22 nations. Few countries maintain records of fires stretching further back in time.

WE CAN SAY LITTLE FOR CERTAIN ABOUT TRENDS IN WILDFIRES WORLDWIDE.

A global clearing house that monitors landscape fires could record the types of vegetation burnt, how severely, over what area and with what loss of life and property. We need these data to facilitate analyses of causes and effects, evaluate fire-management approaches and guide re-insurance rates. They would let us predict how well burning or not burning vegetation might modulate levels of carbon dioxide and other greenhouse gases, and to assess how climate change, land management and socio-economic policies influence fire vulnerability and activity.

Creating a global database will pose a huge policy and technical challenge. But so was building meteorological stations. Some of the elements for such a database already exist in EU and national databases. As well as remote sensing, on-the-ground observations are essential for calibrating and validating these data streams. We could enlist citizen science: smartphones with Global Positioning System access could document biological impacts, and platforms such as Google Earth could help to merge and manage data.

Achieving a global understanding of fire activity will require a major international initiative, and will probably need to be led by an authority such as the World Bank or the United Nations. Although difficult and complicated, the rewards would be invaluable; other projects of this scale reveal the potential: witness the impact of the World Health Organization, and its assessment of the global burden of diseases.

Without improved mapping and monitoring, we will remain unable to answer the most basic questions about trends in wildfires. Acting in ignorance is a poor way to honour — let alone decrease — the lives lost to fires each year. ■

David Bowman *is a pyrogeographer at the University of Tasmania in Hobart, Australia, and studies the biophysical and human dimensions of fire.*
e-mail: david.bowman@utas.edu.au

SEVEN DAYS *The news in brief*

RESEARCH

Ketamine trial

A consumer-advocacy group filed a complaint with the US government on 25 July about two clinical trials in Minnesota that allegedly gave agitated patients ketamine and other sedatives without their consent and despite evidence that doing so could harm their health. Researchers at Hennepin County Medical Center (HCMC) in Minneapolis conducted the trials between 2014 and June 2018. The advocacy group, Public Citizen of Washington DC, submitted the complaint to the US Office for Human Research Protections and the Food and Drug Administration. Sixty-four doctors, bioethicists and academic researchers co-signed the group's letter. A spokesperson for Hennepin Healthcare, which operates the HCMC, declined to comment on the studies until after ongoing internal and external investigations are complete. See go.nature.com/2vajebp for more.



VANUATU RED CROSS VIA EPA/EFE

Volcanic eruption prompts evacuation

The Vanuatu government ordered all 11,000 residents of Ambae island to evacuate on 27 July because of the erupting Manaro Voui volcano. Ash plumes have risen more than 5 kilometres, darkening the sky and making travel difficult as people prepared to move to neighbouring islands. The Vanuatu meteorology and geo-hazards department warned of the health risks of breathing volcanic ash and gas. Manaro Voui is the most voluminous volcano in the South Pacific's New Hebrides archipelago, and an eruption last September forced an island-wide evacuation then, too. It has continued to sporadically belch out ash, triggering several partial evacuations earlier this year.

Embryo editing

Most US adults think that it would be appropriate to use gene editing on human embryos to reduce the risk of developing a serious disease, according to a survey of 2,537 people released on 26 July. The survey, conducted by the Pew Research Center in Washington DC, found that religion had a strong influence on responses. Almost three-quarters of those who reported low levels of religious commitment supported gene editing in embryos to reduce the risk of disease later in life, compared to only 46% of those with high religious commitment. On average, men were more supportive than women of gene editing in embryos, and people with high levels of science knowledge were more comfortable with the idea than were those with less. Only 19% of respondents thought it appropriate to use gene editing to improve intelligence.

Stem-cell advance

Doctors in Japan are poised to implant neural cells made from 'reprogrammed' stem cells into people with Parkinson's disease. It is only the third clinical application of induced pluripotent stem (iPS) cells, which are developed by inducing the cells of body tissues such as skin to revert to an embryonic-like state, from which they can morph into other cell types. Researchers have used the technique to generate precursors to cells that make the neurotransmitter dopamine, which degenerate and die in people with Parkinson's disease. Physicians at Kyoto University Hospital will inject 5 million of these precursor cells into the brains of seven people with the condition. Because dopamine-producing neurons are involved in motor skills, people with the condition typically experience tremors and stiff muscles. Participants will be observed for two years after the transplantation. One of the trial's leaders, stem-cell scientist Jun Takahashi at the Center for iPS Cell Research and Application in Kyoto, demonstrated in 2017 that the therapy improved symptoms in monkeys that had a version of the disease (T. Kikuchi *et al. Nature* **548**, 592–596; 2017). See go.nature.com/2kcgbap for more.

FACILITIES

Telescope boon

A French consortium has become the twelfth member of the Square Kilometre Array (SKA), an effort to build the world's largest radio telescope. The SKA will eventually consist of thousands of radio dishes in Africa and up to 1 million in Australia; construction is expected to begin in 2020. Last year, the coordinating organization scaled back its design for the first phase of the SKA to keep it within a cost cap of €674 million (US$790 million). Project leaders hope that the addition of members, who pay an undisclosed fee, will allow for construction of the original design. Spain joined the effort in June. The Maison

© 2018 Springer Nature Limited. All rights reserved.

SKA-France consortium is the first public–private group to join the SKA and is headed by the country's largest research agency, the CNRS. The consortium includes four other research organizations and seven private companies. The SKA's other members comprise 11 nation states, including China and India.

HEALTH

Ebola outbreak ends

An outbreak of the Ebola virus that began in the Democratic Republic of the Congo (DRC) in early May was declared over by the country's health ministry on 24 July. A total of 54 people were infected, 33 of whom died. The outbreak was centred in remote rural regions of the northwestern Équateur Province, but cases were also reported in Mbandaka, a city on the banks of the Congo River, a major transport route. That prompted fears that the disease could spread quickly both in the city and to other regions. The outbreak is the DRC's ninth; the first known outbreak of Ebola virus occurred in the nation in 1976. Learning from mistakes in the large 2014–16 Ebola epidemic in West Africa, international public-health bodies quickly gave the DRC



JUNIOR D. KANNAH/AFP/GETTY

support — including helping to deploy several thousand doses of an experimental Ebola vaccine (**pictured**).

Fetal trial stopped

A clinical trial in the Netherlands designed to test whether sildenafil (Viagra) could treat a fetal growth condition has been cancelled after an interim assessment found that the therapy caused harm. The randomized trial treated 90 pregnant women with sildenafil and 93 with a placebo. Seventeen fetuses in the treatment group developed a severe lung disease and 11 of them died from the disease; only 3 fetuses in the control group developed the lung disease, and none of them died (fetuses in both groups also died from causes other than the lung disease). The growth condition is caused by insufficient blood oxygen and nutrient supply through the placenta, and carries a high risk of fetal death or of brain damage in surviving babies. Researchers had hoped that sildenafil would dilate blood vessels in the placenta and improve fetal growth. The scientists running the trial notified Canadian researchers conducting a similar study, who have temporarily stopped their research.

PEOPLE

University lawsuit

The latest in a string of lawsuits filed against the University of Southern California (USC) in Los Angeles claims that the university "suppressed and concealed" years of sexual-harassment complaints against former gynaecologist George Tyndall. The complaints from former and current students span nearly three decades, and include allegations that Tyndall "perpetrated serial sexual abuse, harassment, molestation, and violation", according to one of the lawsuits filed on 23 July. After an internal investigation in 2016, the USC suspended Tyndall with pay, and he retired in 2017. California's Department of Education is currently investigating the university's response to the complaints. In a statement to *Nature*, the USC says that it will seek a "prompt and fair resolution that is respectful of our former students". *Nature*'s attempts to reach Tyndall and his lawyer were unsuccessful.

POLICY

Trial reporting

The US National Institutes of Health (NIH) has postponed enforcement of a controversial rule that would have required behavioural studies to register as clinical trials. In a 20 July announcement, the agency said that until 24 September 2019, researchers performing basic trials with human participants would not be penalized for failing to list their studies on the government database ClinicalTrials.gov. Brain and behavioural scientists had pushed back against clinical-trial reporting policy that the NIH announced in 2016, which held non-invasive studies — such as cognitive tests and brain scans — to the same standards as drug trials. Numerous researchers argued that these rules were too burdensome and confusing. In its latest announcement, the NIH asked scientists to submit suggestions of how to make the policy more appropriate for basic research.

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

TREND WATCH

Only 13% of Earth's oceans can now be classed as wilderness, according to a 26 July study (K. R. Jones *et al. Curr. Biol.* https://doi.org/csh5; 2018).

Researchers used global data on 15 human stressors of oceans, including fishing, pollution and commercial shipping. Areas were defined as wilderness if they showed little impact from these stressors (scoring in the bottom 10% of a measure of each), as well as a low aggregate score that combined these human activities and climate stresses such as ocean acidification.

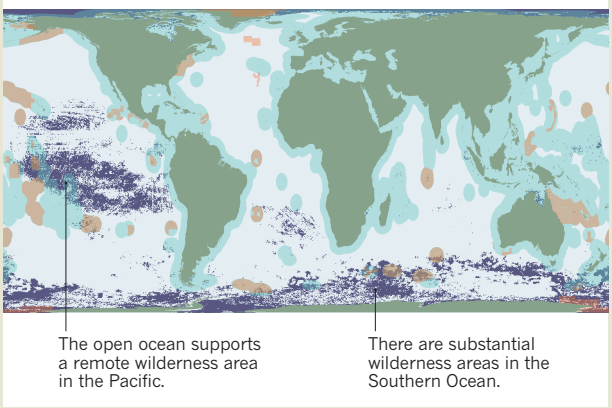
Most wilderness areas are in the open ocean and around the poles, far from human populations. Coastal ecosystems — which include centres of biodiversity such as coral reefs — make up just 10% of the wilderness area. Of all marine wilderness, just 5% is covered by marine protected areas.

"High seas at the moment are a Wild West," says author Kendall Jones, a conservation biologist at the University of Queensland in Brisbane, Australia. "It really highlights that those last few areas that are not impacted — how important they are."

The United Nations is debating a high-seas conservation treaty, which should be signed by 2020.

WILD OCEANS

Just 13% of Earth's oceans should be considered wilderness, according to a study of global impacts on the seas. Marine protected areas encompass just 5% of the remaining wilderness.



SOURCE: K. R. JONES *ET AL. CURR. BIOL.* HTTPS://DOI.ORG/CSH5 (2018)

© 2018 Springer Nature Limited. All rights reserved.

NEWS IN FOCUS

CHINA Vaccine scandal won't put parents off immunizing their children **p.14**

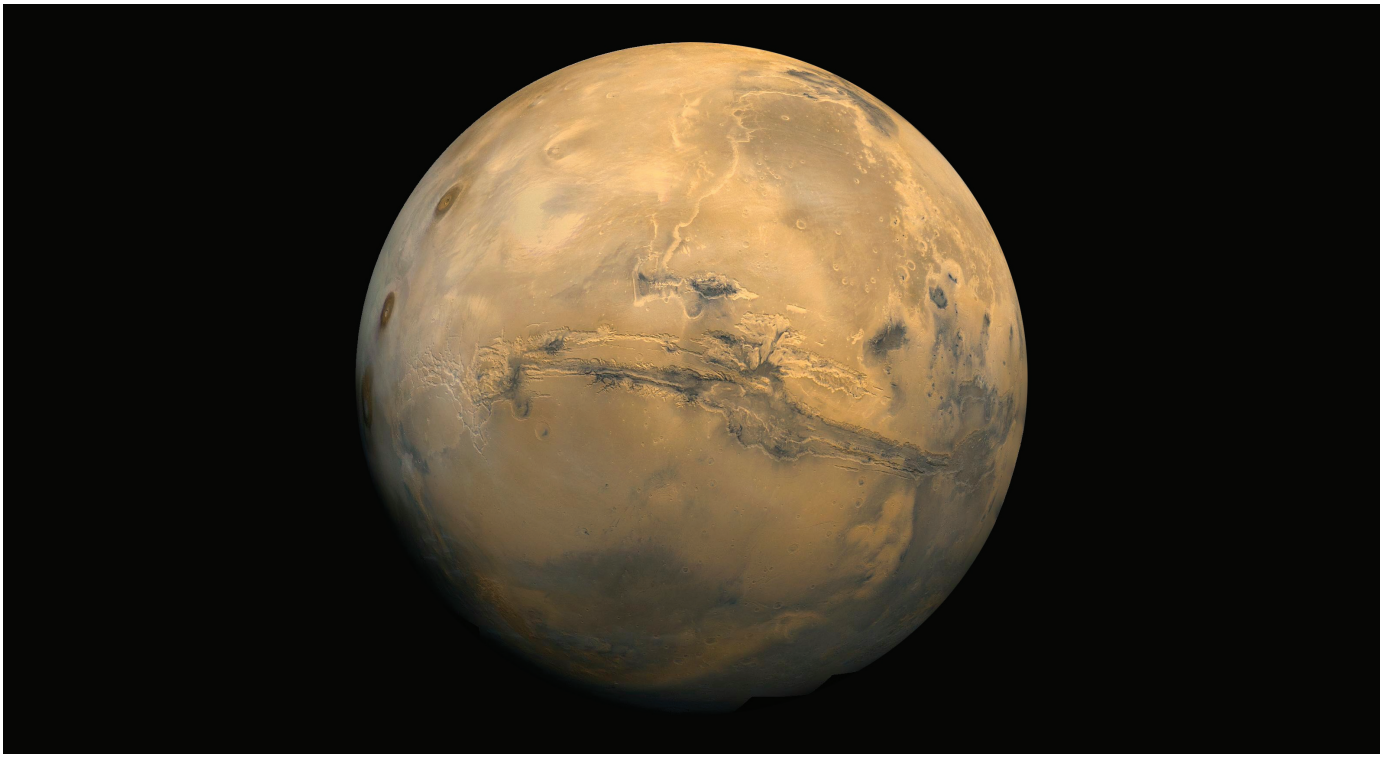
POLICY Scientists disappointed by European court ruling on gene-edited plants **p.16**

CONSERVATION Efforts to gut US Endangered Species Act face resistance **p.17**



HEAT ADVISORY Assessing the role of climate change in natural disasters **p.20**

NASA



Mars is thought to host a buried lake that could change how scientists want to explore the red planet.

PLANETARY SCIENCE

Signs of buried lake on Mars tantalize scientists

If confirmed, the lake would be the first body of liquid water ever detected on the red planet.

BY ALEXANDRA WITZE

A large saltwater lake seems to lurk under ice near Mars's south pole. If confirmed, it would be the first body of liquid water ever detected on the red planet, and a significant milestone in the quest to determine whether life exists there.

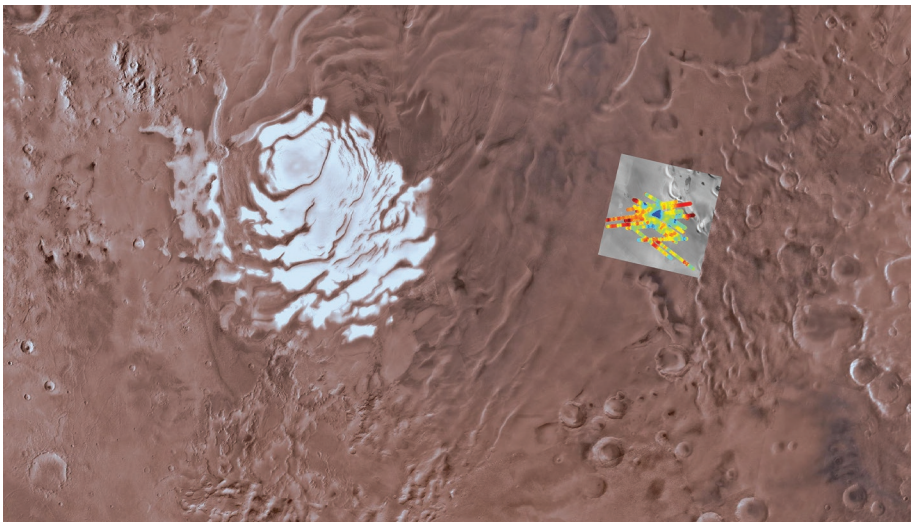
"It's a very promising place to look for life on Mars," says Roberto Orosei, a planetary scientist at the National Institute of Astrophysics in Bologna, Italy. "But we do not know for sure if it is inhabited." On Earth, similar 'subglacial' lakes are home to microbial life.

A team of Italian researchers, led by Orosei, reported the discovery on 25 July in *Science*[1]. They spotted evidence of the buried lake in radar data from the European Space Agency's Mars Express spacecraft.

Others say that the work is tantalizing but, like anything else in the controversial hunt for water on Mars, it needs more supporting evidence. "It's not quite a slam dunk yet," says Jeffrey Plaut, a planetary scientist at NASA's Jet Propulsion Laboratory in Pasadena, California, who has searched for water using data from Mars Express[2].

If further studies confirm the existence of a lake, the discovery could open new avenues for investigating Mars. Researchers have drilled into subglacial lakes on Earth and sampled the water for signs of microbes, while others are developing technologies to reach a buried ocean on Jupiter's moon Europa. There are no ice-drilling missions currently slated for Mars — but the latest discovery could change how scientists think about exploring the planet.

"It begins a new line of inquiry that's very exciting," says Jim Green, NASA's chief scientist. Water appears across Mars today in various forms, left over from a time billions of years ago when the planet was warmer and wetter. ▶



USGS ASTROGEOLOGY SCIENCE CENTER, ARIZONA STATE UNIVERSITY, INAF

Radar tracks on Mars's Planum Australe show the location of a potential buried lake (in blue).

Orbiting probes have spotted ice, including buried glaciers, in many locations. Spacecraft have photographed steep slopes whose appearance changes seasonally, as if liquid water is running downhill and leaving dark marks. And NASA's Curiosity rover has measured water vapour in the planet's atmosphere.

Orosei and his colleagues found the lake using a radar instrument called MARSIS aboard Mars Express, which launched in 2003. It sends radio waves bouncing off the planet's surface and subsurface layers; the way in which the radar signal reflects back reveals the type of material that is present, such as rock, ice or water. The scientists focused their search on the layers of ice and dust that cover the planet's south pole.

But the observations were frustratingly inconsistent. Mars Express sometimes saw a bright reflection in several locations, which did not reappear the next time the spacecraft flew over those sites. Finally, in 2012, the scientists decided to get MARSIS to send back raw data, instead of performing automated processing before beaming the data to Earth. "This changed everything, and it was much more obvious to spot the bright reflectors," says Orosei.

The data showed the reflections coming from a 20-kilometre-long zone in a region known as Planum Australe. After ruling out other possible causes, such as carbon dioxide ice, the scientists concluded that the reflections were coming from subsurface water.

The lake is about 1.5 kilometres beneath Mars's icy surface and is at least 1 metre deep. To keep from freezing, the water must be very salty, Orosei says — perhaps similar to super-salty subglacial lakes reported in the Canadian Arctic earlier this year[3]. Salt-rich rocks beneath the Canadian lakes infuse the water and allow it to remain liquid, says Anja Rutishauser, a glaciologist at the University of Alberta in Edmonton. On Mars, salts known as perchlorates might be making the brine; in 2008, NASA's Phoenix spacecraft found perchlorates in soils near the planet's northern polar ice.

Mars might have had many similar lakes in the past, when heat rising from deep in the planet melted some of the ice covering its polar regions, says Stephen Clifford, a planetary scientist who proposed the idea[4] in 1987 and now works for the Planetary Science Institute in Houston, Texas. If life once thrived in ancient subsurface lakes, he says, the latest finding "raises support for the idea that life could still persist on Mars".

With liquid water and the right chemical elements available to supply energy, a buried martian lake would have the ingredients needed to sustain life — as long as it's not too salty, says John Priscu, a biogeochemist at Montana State University in Bozeman. But exploring it won't be easy. Priscu leads a team that aims to drill into Antarctica's subglacial Lake Mercer later this year; hauling tonnes of equipment and fuel there required weeks of traversing the Antarctic ice sheet with tractors. "There's no way you're going to get all that to Mars," he says.

But there are ways to learn more with spacecraft already in play. Green notes that NASA's InSight probe, which is scheduled to land near the martian equator in November, will measure heat flow in the top 5 metres of the surface there. Scientists can use those data to extrapolate how much heat might be rising from beneath the south polar cap, melting the ice and potentially creating more lakes.

Orosei says his team has glimpsed other bright reflections, but isn't ready to say whether or not they are lakes. More studies using MARSIS, as well as the radar on board NASA's Mars Reconnaissance Orbiter — which has looked at Planum Australe and not seen the reflections — could help to reveal whether these are actually liquid water or something else, Plaut says. ■

1. Orosei, R. *et al. Science* https://doi.org/10.1126/science.aar7268 (2018).
2. Plaut, J. J. *et al. Science* **316**, 92–95 (2007).
3. Rutishauser, A. *et al. Sci. Adv.* **4**, eaar4353 (2018).
4. Clifford, S. M. J. *Geophys. Res.* **92**, 9135–9152 (1987).

PUBLIC HEALTH

China vaccine scandal unlikely to dent immunization rates

Vaccines are compulsory for children starting school in China, and enjoy public support.

BY NICKY PHILLIPS

Problems with two Chinese-made vaccines — one of which was distributed to clinics and possibly injected into hundreds of thousands of children — have led to arrests and made international headlines. But researchers who study vaccination in China don't expect a major effect on the country's high immunization rates.

Widespread support for immunization programmes, combined with strict vaccine requirements for children starting school, means that most parents will continue vaccinating their children, they say.

"I don't think there'll be an appreciable drop in vaccine coverage but it could impact when people get vaccines, and where the vaccines come from," says Abram Wagner, an epidemiologist at the University of Michigan in Ann Arbor who has interviewed Chinese parents

about their views on immunization.

On 15 July, China's national drug watchdog revealed that during a surprise inspection of vaccine maker Changchun Changsheng Biotechnology's facilities in the province of Jilin, it found that the company had faked production data for several batches of the rabies vaccine. Authorities ordered that the doses be disposed of and revoked the company's manufacturing permit for that vaccine; it is not clear whether anyone received the faulty doses.

Five days later, local government investigators announced that the same company had violated standards in making about 250,000 doses of the DTP vaccine, which protects against diphtheria, tetanus and pertussis (whooping cough), rendering the doses potentially ineffective. For that breach, the company says it was fined 3.44 million yuan (US$505,000).

It is not known how many children received the faulty DTP vaccines, which were recalled when the problem was uncovered by authorities in November, but so far no health issues have been reported. The main concern is that these vaccines won't protect children from the dangerous infections that they're meant to combat, says Wagner.

Parents turned to social media to voice their anger at the company and their concerns about domestic vaccines. But Xiaomin Wang, a social scientist at Zhejiang University Institute for Social Medicine in Hangzhou, agrees that the scandal is unlikely to reduce child immunization rates.



VCG/GETTY

China has one of the world's highest rates of vaccination coverage.

PARENTS SURVEYED

In 2016, when Chinese authorities discovered that childhood vaccines rendered ineffective by improper storage had been distributed to medical clinics across the country over five years, Wang and her colleagues went out and asked parents about their views on vaccines. They found that parents had very low faith in the safety of domestically produced vaccines, which make up 95% of vaccines given in China — only 11% said they trusted them. But the researchers also found that more than half of parents surveyed still intended to rely on them to vaccinate their children (X. Wang *et al. J. Health Commun.* **23**, 413–421; 2018).

The disparity between parents' trust in domestic vaccines and their willingness to use them is probably influenced by several factors, including cost and availability, says one of Wang's collaborators on the survey, Leesa Lin, who studies social behaviour and risk perception at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts.

Although the government subsidizes many domestic-made vaccines, those who have access to and can afford foreign-made vaccines are likely to seek them out, says Lin. "When an incident like this happens, the public might seek safer alternatives but would not give up on vaccination," she says.

She attributes this, in part, to a widespread understanding of the benefits of vaccines, after decades of government campaigns promoting immunization. China has one of the world's highest vaccination coverage rates — for example, 99% of Chinese infants receive the required three doses of the DTP vaccine, compared with 85% of infants globally, and 95% in the United States. Lin and the team are about to submit for publication results from another survey, which confirmed parents' strongly-held faith in the benefits of vaccination programmes.

Those results stand in contrast to the attitudes of some small groups of parents in Europe and the United States who resist vaccinating their children, citing unfounded safety concerns or religious reasons, says Lin. "In China, those factors do not play much of a role."

Lin and Wang plan to survey parents again in the coming months to improve understanding of the factors that influence their decisions.

Wagner notes that the DTP vaccine is mandatory for children starting school in China and that such requirements contribute to the country's high vaccine uptake. Some parents might delay vaccination in the wake of the latest scare, but they are unlikely to risk their child being denied entrance to school, he says. He adds that it's particularly hard to get an exemption from such requirements in China, compared with other countries that have similar rules.

REPAIR TRUST

Wagner hopes that Chinese officials will be more transparent about what happened at Changchun Changsheng than they were during the 2016 scandal, and that they will take stronger action to prevent another incident. At the time, the government promised to improve oversight of vaccine manufacturing and transportation, he says. "They talked big, but I'm not exactly sure what they did," he says. "I hope that they'll learn some lessons from this event and implement tighter regulations."

There are already encouraging signs. Several Changchun Changsheng executives, including the chair, have been arrested, and Chinese President Xi Jinping has said that the events are shocking and has ordered an investigation of the vaccine production chain. "This is a step forward compared to last time," says Lin.

China's vaccine makers will also have to convince international markets that their vaccines are safe and effective if the country is to become a major global producer, says Wagner. "Vaccines and pharmaceutical products could be a huge industry for them." ■

MORE ONLINE

STAY CURRENT

● Yeast cell's genome squeezed onto a single chromosome
go.nature.com/2vkl2mt
● Controversial trial that gave ketamine without consent sparks ethics complaint
go.nature.com/2vajebp

NATURE PODCAST

How a bird sees colour; modifying Mars; and linking extreme weather to the changing climate
nature.com/nature/podcast



CHRIS RATCLIFFE/BLOOMBERG/GETTY

In the EU, gene-edited crops and food will be treated in the same way as genetically modified organisms.

GENE EDITING

EU law deals blow to CRISPR crops

Top court's ruling threatens research on gene-edited plants.

BY EWEN CALLAWAY

Gene-edited crops should be subject to the same stringent regulations that govern conventional genetically modified (GM) organisms, Europe's highest court ruled on 25 July.

The decision, handed down by the Court of Justice of the European Union (ECJ) in Luxembourg, is a major setback for proponents of gene-edited crops, including many scientists. They had hoped that organisms created using relatively new, precise gene-editing technologies such as CRISPR–Cas9 would be exempted from existing European law, which has limited the planting and sale of GM crops.

Instead, the ECJ ruled that crops created using these technologies are subject to a 2001 directive. That law was developed for older breeding techniques, and it puts high hurdles in the way of developing GM crops for food.

"It is an important judgment, and it's a very rigid judgment," says Kai Purnhagen, a legal scholar at Wageningen University and Research in the Netherlands who specializes in European and international law. "It means for all the new inventions, such as CRISPR–Cas9 food, you would need to go through the lengthy approval process of the European Union."

That is likely to hinder investment in crop research using these tools in the EU, says Purnhagen. "From a practical perspective, I don't think this will be at all of interest for business. So they will move somewhere else," he says.

The ruling is "tremendously disappointing", says Nigel Halford, a crop geneticist at Rothamsted Research in Harpenden, UK. "It's a real hit to the head," he says. Gene-editing techniques will still be used as a research tool for developing crops, he adds, but he doubts that companies in Europe will have much appetite to develop them. "They are not going to invest in a technology they see not having any commercial application," Halford says.

Environmental organization Friends of the Earth in Amsterdam, meanwhile, applauded the court's decision in a statement. It also called for all products made through gene editing to be regulated, assessed for their health and environmental impacts, and labelled.

DNA CHANGES

The 2001 EU directive behind the ECJ's decision concerns the intentional release of GM organisms into the environment — and was aimed at species into which entire genes, or long stretches of DNA, had been inserted. The law exempts organisms whose genomes were modified using 'mutagenesis' techniques, such as irradiation, which introduce changes to an organism's DNA but don't add foreign genetic material.

In 2016, the French government asked the ECJ to interpret the directive in light of plant-breeding techniques that have since emerged.

Many plant breeders and scientists contend that gene-editing techniques such as CRISPR–Cas9 should be considered mutagenesis, just like irradiation, and thus be exempt from the directive, because they can involve changes to DNA and not the insertion of foreign genes. But people opposed to GM organisms contend that the deliberate nature of alterations made through gene editing means that they should fall under the directive.

In January, an advocate-general with the court, Michal Bobek, issued a 15,000-word opinion that both sides claimed was partly in their favour. He said that gene-edited crops do constitute GM organisms under the original directive, but also that species modified using technologies discovered since 2001 — such as those used for gene editing — could be exempted, as long as they don't contain DNA from other species, or artificial DNA.

But in its ruling, the ECJ determined that only mutagenesis techniques that have "conventionally been used in a number of applications and have a long safety record are exempt from those obligations". Organisms made using mutagenesis techniques developed after 2001 — including gene editing — are not exempt from the directive.

NO INCENTIVE

"This will have a chilling effect on research, in the same way that GMO legislation has had a chilling effect for 15 years now," says Stefan Jansson, a plant physiologist at Umeå University in Sweden. Gene-edited crops will not vanish from European research labs, but he worries that the funding to develop them could dry up. "If we cannot produce things that society finds helpful, then they will be less likely to fund us."

Jansson also has practical concerns about the ruling. He developed a 'CRISPR cabbage' that he has consumed, and which was growing in his home garden as he spoke to *Nature*. "I took a photo yesterday, and I took another after the ruling. It's still the same plant. Yesterday it wasn't a GMO, and now it's a GMO. I'm a bit curious what I have to do. Do I have to remove it?"

Purnhagen says that the ruling leaves open a possible loophole, whereby if scientists can prove that gene-editing techniques are as safe as mutagenesis methods already exempt from the law, such as irradiation, the new techniques, too, could earn an exemption.

But he doubts that researchers and businesses developing gene-edited crops will hold out hope. "I can't see CRISPR–Cas9 and all these new technologies will be profitable in the European Union. I can't see this happening. I think this research will move somewhere else." ■

ASTROPHYSICS

Black hole provides test for Einstein's theory

General relativity seen in action at Milky Way's centre.

BY ALEXANDRA WITZE

Astronomers have caught the giant black hole at the Galaxy's centre stretching the light emitted by an orbiting star — nearly three decades after they first started tracking the star. The long-sought phenomenon, known as gravitational redshift, was predicted by Albert Einstein's general theory of relativity, but until now, it had never been detected in the environs of a black hole.

"It's another big step in getting closer to understanding the black hole," says Heino Falcke, an astronomer at Radboud University in Nijmegen, the Netherlands, who was not involved in the research. "This is just amazing, to be able to see these effects."

A team led by astrophysicist Reinhard Genzel of the Max Planck Institute for Extraterrestrial Physics in Garching, Germany, reported the discovery on 26 July in *Astronomy & Astrophysics* (R. Abuter *et al. Astron. Astrophys.* **615**, L15; 2018).

Genzel and his colleagues have been tracking the journey of this star, known as S2, since the early 1990s. Using telescopes at the European Southern Observatory in Chile, the scientists watch as it travels in an elliptical orbit around the black hole, 8,000 parsecs (26,000 light years) from Earth. With a mass 4 million times that of the Sun, the black hole generates the strongest gravitational field in the Milky Way. That makes it an ideal place to hunt for relativistic effects.

On 19 May, S2 passed as close as it ever does to the black hole. The researchers traced the star's path using instruments including GRAVITY, an interferometer that combines light from four 8-metre telescopes. GRAVITY measured S2's movement across the sky; at its fastest, the star whizzed along at more than 7,600 kilometres per second, nearly 3% of the speed of light. A different instrument studied how fast S2 moved towards and away from Earth as the star swung past the black hole. Combining the observations let Genzel's team detect S2's gravitational redshift — which describes how its light is stretched by the black hole's immense gravitational pull. Such a phenomenon is consistent with the predictions of general relativity.

Future observations of S2 might confirm other predictions made by Einstein, such as how a spinning black hole drags space-time around with it.

"Their data look beautiful," says Andrea Ghez, an astronomer at the University of California, Los Angeles, who leads a competing team that uses the Keck telescopes in Hawaii to measure the star's path around the Galactic Centre. It takes 16 years for S2 to make a complete orbit around the black hole, so both groups have been eagerly awaiting this year's close passage. They are still watching the star closely as it slows to its minimum velocity in the line of sight from Earth — another crucial event. "We're in the thick of it," says Ghez. "It's super-exciting." ■

POLICY

US wildlife law in danger

But bills that could strip protections from vulnerable species face resistance from legislators.

BY JEREMY REHM

The US Endangered Species Act — which protects more than 2,000 species of plants and animals, including insects, at risk of extinction — is under renewed attack from Republican politicians. But policy experts say that these efforts face an uphill battle, even though Republicans control the White House and both chambers of Congress.

On 19 July, the US Fish and Wildlife Service (FWS) and the National Marine Fisheries Service (NMFS) proposed policy changes that would, among other things, make it easier to delist species and harder to add new ones. And in recent weeks, legislators in the US House of Representatives have gone further by introducing about 12 bills aimed at altering the law itself.

Some of the bills would roll back protections for species including the American burying beetle (*Nicrophorus americanus*). Lawmakers say that this is to remove barriers to the activities of businesses such as oil companies. Other bills propose fundamental changes to the law, for example by narrowing the range of habitats deemed necessary for organisms to recover.

"The law needs to be updated to ensure it maintains its original intent and focus of species recovery, and not simply serve as a tool for endless litigation," says Representative Rob Bishop, the Utah Republican who heads the House of Representatives Committee on Natural Resources. The committee has spearheaded many of the bills under consideration.

"It's pretty much the oil and gas companies that have been looking to have the beetle delisted."

Decisions on whether or not to list a species under the Endangered Species Act (ESA) often draw legal challenges from industry and environmental organizations. Special-interest groups including the oil and gas industry have fought the law since its enactment in the 1970s.

Many conservation scientists and environmental groups say the Republican legislation would cripple the ESA by making it much harder to protect species that are now imperilled. One of the House bills would weaken safeguards for threatened species.

Brett Hartl, government-affairs director for the Center for Biological Diversity in Washington DC, doubts that the bills will become law — even if the Republicans retain control of the US Senate and the House after the November midterm elections. Similar legislation that has been introduced over the past several years has foundered, he notes.

Hartl worries more about the changes to the ESA proposed in mid-July by the FWS and the NMFS. "Those are very dangerous," he says, because they don't require approval by Congress to take effect. The plans must undergo a 60-day public-comment period before President Donald Trump's administration can finalize the changes and implement them. Hartl thinks that they will probably make it through the process.

But the plan is likely to face lawsuits, says Steve Holmer, vice-president of policy at the American Bird Conservancy in The Plains, Virginia. Litigants could challenge the proposals on several grounds, using arguments that the changes would harm species — in direct opposition to the ESA's original aim.

One of the species that researchers worry about is the endangered American burying ▶



JOEL SARTORE/GETTY

The endangered American burying beetle inhabits areas that are of interest to oil and gas companies.

beetle, which legislators have been trying to strip of federal protections since 2013.

Habitat destruction in the twentieth century eliminated 90% of this insect's historical range, which stretched across 35 states in the Midwest and the East Coast. Declining food sources also contributed to population declines. The US government added the beetle to the endangered-species list in 1989.

The insect's remaining habitat includes parts of Kansas, Oklahoma and Nebraska — states where gas- and oil-drilling companies hold interests, says Louis Perrotti, director of conservation programmes at Roger Williams Park Zoo in Providence, Rhode Island. "It's pretty much the gas and oil companies that have been looking to have the beetle delisted."

The most recent legislative salvo included an addition to the House's 2019 funding bill for the Department of Defense that would have removed the insect from the endangered-species list. After an outcry from some of the lawmakers working to reconcile the House and Senate versions of the spending bill, legislators removed the addition early last week.

Perrotti and his collaborators have been breeding burying beetles in captivity and releasing them into the wild to establish self-sustaining populations. The insects are important because they feed on carrion. "Without burying beetles, we'd be knee-deep in dead and decaying carcasses," says Perrotti.

If lawmakers eventually succeed in delisting the insect, the repopulation project could lose major collaborators that receive federal funding. And there is a real risk of the beetle going extinct if legislators change the ESA, says Perrotti. Nevertheless, he and his collaborators will keep pushing to save the species. "A lot of people have put their blood and soul into this, including me." ■

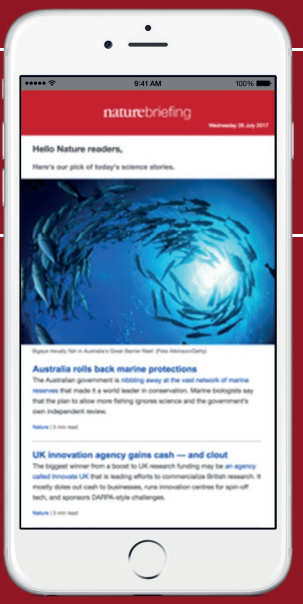
CLARIFICATION

The World View 'Preprints could promote confusion and distortion' (*Nature* **559**, 445; 2018) omitted to mention that funders of the Science Media Centre (the author's employer) include Springer Nature, *Nature*'s publisher.

nature briefing

What matters in science and why – free in your inbox every weekday.

The best from *Nature's* journalists and other publications worldwide. Always balanced, never oversimplified, and crafted with the scientific community in mind.



SIGN UP NOW
go.nature.com/briefing

nature

A45829



NATALIA KOLESNIKOVA/AFP/GETTY

Climate as culprit

Weather forecasters will soon provide instant assessments of global warming's influence on heatwaves and floods.

BY QUIRIN SCHIERMEIER

The Northern Hemisphere is sweating through another unusually hot summer. Japan has declared its record temperatures a natural disaster. Europe is baking under prolonged heat, with destructive wildfires in Greece and, unusually, the Arctic. And drought-fuelled wildfires are spreading in the western United States.

For Friederike Otto, a climate modeller at the University of Oxford, UK, the past week has been a frenzy, as journalists clamoured for her views on climate change's role in the summer heat. "It's been mad," she says. The usual scientific response is that severe heatwaves will become more frequent because of global warming. But Otto and her colleagues wanted to answer a more particular question: how had climate change influenced this specific heatwave? After three days' work with computer models, they announced on 27 July that their preliminary analysis for northern Europe suggests that climate change made the heatwave more than twice as likely to occur in many places.

Soon, journalists might be able to get this kind of quick-fire analysis routinely from weather agencies, rather than on an ad hoc basis from academics. With Otto's help, Germany's national weather agency is preparing to be the first in the world to offer rapid assessments of global warming's connection to particular meteorological events. By 2019 or 2020, the agency hopes to post its findings on social media almost instantly, with full public reports following one or two weeks after an event. "We want to quantify the influence of climate change on any atmospheric conditions that might bring extreme weather to Germany or central Europe," says Paul Becker, vice-president of the weather agency, which is based in Offenbach. "The science is ripe to start doing it".

The European Union is interested too. The European Centre for Medium-Range Weather Forecasts (ECMWF) in Reading, UK, is preparing to pilot a similar programme by 2020 that will seek to attribute

Global warming tripled the odds of Russia's extreme heatwave in 2010, which exacerbated wildfires.

extreme events, such as heatwaves or floods, to human-induced climate change. If that works well, a regular EU attribution service could be in place a year or two later, says Richard Dee, head of the EU's Copernicus Climate Change Service at the ECMWF. "It's ambitious, but doable," says Otto, who is also helping to set up the EU effort.

That weather agencies are contemplating such regular services shows how far 'attribution science' has come since the first cutting-edge research projects — more than a decade ago — tried to attribute individual weather events to climate change[1]. Now, after more than 170 studies in peer-reviewed journals, attribution science is poised to burst out of the lab and move into the everyday world[2]. It still has difficulty with some kinds of extreme weather phenomena, but as meteorological services begin to offer attribution information routinely, the bigger challenge is to work out how to make the studies helpful to the people who might use them. "It's one thing to make scientifically robust attribution statements," says Peter Walton, a social scientist at the University of Oxford. "How to go about using that information is another thing."

ATTRIBUTION 101

The idea behind attribution science is simple enough. Disasters such as record-breaking heatwaves and extreme rainfall are likely to become more common because the build-up of greenhouse gases is altering the atmosphere. Warmer air contains more water vapour and stores more energy; the increasing temperatures can also change large-scale atmospheric circulation patterns. But extreme weather can also arise from natural cycles, such as the El Niño phenomenon that periodically warms sea surface temperatures in the tropical Pacific Ocean.

Researchers say that teasing out the role of human-induced global warming — as opposed to natural fluctuations — in individual weather extremes will help city planners, engineers and home-owners to understand which kinds of floods, droughts and other weather calamities are increasing in risk. And surveys suggest that people are more likely to support policies focused on adapting to climate-change impacts when they have just experienced extreme weather, so quickly verifying a connection between a regional event and climate change, or ruling it out, could be particularly effective[3].

Otto, the deputy director of the University of Oxford's Environmental Change Institute, is a veteran of attribution science, having conducted more than two dozen analyses. On 4 June, for instance, she and her colleagues completed a study focused on the southern edge of Africa, which had been suffering from a three-year drought. By early this year, the situation had become so dire in South Africa's Western Cape Province that officials in Cape Town had warned they would soon hit 'Day Zero', when the region would run out of water to serve basic needs — a first for a major city.

As reports of Day Zero made international headlines, Otto and Mark New, a climate scientist at the University of Cape Town, decided that the event was a good candidate for an attribution study. Working in their spare time because they had no dedicated funding for the project, researchers from the Netherlands, South Africa, the United States and the United Kingdom started by defining the regional extent of the multi-year drought. They also created an index of its severity, which combined measurements of rainfall and heat. Then, the teams turned to the workhorses of attribution studies: complex computer models that mimic Earth's climate. On each of five independent models, they ran thousands of simulations. Some of these took into account observed levels of human-generated greenhouse gases; others ran with natural concentrations of the gases, as if the Industrial Revolution had never happened. The researchers compared how many times a drought of similar severity and extent turned up in the thousands of test runs. Most of the teams used their own dedicated computers, but the Oxford branch of the study conducted its simulations on the weather@home model ensemble, a distributed computing framework that uses the idle time of thousands of volunteers' personal computers.

By the time the team met in June, rains had returned to South Africa and had pushed Day Zero away. But the scientists were still chasing the causes of the mega-drought, which could help to determine whether the region might face a repeat anytime soon. Coordinating a four-way Skype call from her office in Oxford, Otto looked relieved when colleagues agreed that the analysis had yielded a result. "Global warming has tripled the odds of three consecutive dry years in the region," she says.

The findings came just in time for Roop Singh, a climate-risk adviser at the Red Cross Red Crescent Climate Centre in The Hague, the Netherlands, to present the results at a conference on climate-change adaptation in Cape Town two weeks later. Researchers there didn't find the results particularly shocking, Singh says — but they did trigger lively discussions about whether the increase in drought risk could help to justify increased investment in diversifying water sources in Cape Town. Otto's study was published on 13 July, before peer review, at the website of World Weather Attribution (see go.nature.com/2tyjezc), a partnership of six research institutes (including the University of Oxford) that joined together in 2014 to analyse and communicate the possible effect of climate change on extreme weather events.

Although Cape Town avoided Day Zero this year, policymakers in the region say Otto's results send a sobering warning to water authorities that might be inclined to downplay the risk of global warming. "This is an incredibly strong message which we cannot afford to ignore," says Helen Davies, director of green economy in the Western Cape Government's Department of Economic Development and Tourism. "We may need to work on a radically new approach to water management," she says.

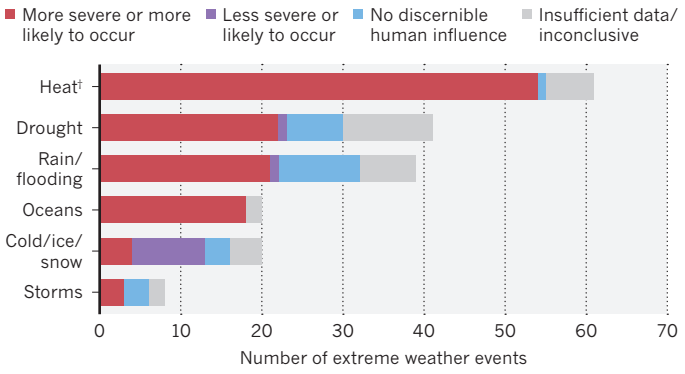
The work by Otto's team joins a rapidly growing corpus of studies on climate attribution. From 2004 to mid-2018, scientists published more than 170 reports covering 190 extreme weather events around the world, according to an analysis by *Nature*, which builds on previous work by the publication CarbonBrief (see go.nature.com/2jypsyc). So far, the findings suggest that around two-thirds of extreme weather events studied were made more likely, or more severe, by human-induced climate change (see 'Attribution science'). Heat extremes made up more than 43% of these kinds of events, followed by droughts (18%) and extreme rain or flooding (17%). In 2017, for the first time, studies even stated that three extreme events would not have occurred without climate change: heatwaves in Asia[4] in 2016, global record heat in the same year[5], and marine warming in the Gulf of Alaska and the Bering Sea[6] from 2014–16. But in 29% of cases in *Nature*'s analysis, the available evidence either showed no clear human influence or was too inconclusive for scientists to make any judgement.

> *"If we want science to be part of the discussion, we need to say something fast."*

Sometimes studies seem to come to opposite conclusions about a particular event. One study about a 2010 heatwave in Russia found that its severity was still within the bounds of natural variability[7]; another analysis determined that climate change had made the event more likely to occur[8]. The media found the results confusing, but climate scientists say the discrepancy is not surprising because the two studies looked at different issues: severity and frequency[9]. According to Otto, "The example goes to show that framing and communicating attribution questions is a real challenge." But researchers have become more sophisticated since then about how they set up and present their studies, she adds.

Attribution science

Researchers have published more than 170 studies* examining the role of human-induced climate change in 190 extreme weather events.



*Studies from 2004–18 collated by *Nature* and CarbonBrief. †Heat includes heatwaves and wildfires; Oceans includes studies on marine heat, coral bleaching and marine-ecosystem disruption.

RAPID REPORTS

The South Africa study could have been done faster, had the researchers been able to spend all their time on it. This year's work during the European heatwave was not the first rapid study: in 2015, for instance, during another sweltering heatwave in Europe, an international team of researchers (including Otto) found within weeks how climate change had made comparable heatwaves four times more likely in some European cities, and at least twice as likely over much of the continent (see go.nature.com/2uzpoxc). Meteorological agencies plan to work even faster when they put these experimental methods into regular operation. Over the past few months, Otto has talked extensively with the staff of the German weather service, briefing them on how to conduct attribution studies using the best approaches. On 21 June, she signed an agreement with the agency that provides free use of the University of Oxford's weather@home model. Meanwhile, the Copernicus Climate Change Service has asked Otto and two of her colleagues to write a paper describing workflows and methods for conducting rapid attribution studies, to be published by September.

Otto says a rapid attribution service is needed because questions about the role of climate change are regularly asked in the immediate aftermath of extreme weather events. "If we scientists don't say anything, other people will answer that question not based on scientific evidence, but on whatever their agenda is. So if we want science to be part of the discussion that is happening, we need to say something fast," she says.

Some scientists might feel uncomfortable if weather forecasters announce results before work has gone through peer review. But in these cases, the methods have already been extensively reviewed, says Gabriele Hegerl, a climate scientist at the University of Edinburgh, UK. Hegerl is also a co-author of a 2016 report by the US National Academies, which concluded that the science of attribution has advanced rapidly and would benefit from being linked to operational weather prediction. "It can be really useful to have results quickly available for event types we understand reasonably well, such as heatwaves," she says. "You don't need to peer review the weather forecast," adds Otto.

But not all of the science involved in attribution studies is settled, Hegerl says. Computer algorithms still struggle to model severe local storms that result from the rapid convection of air, such as small hailstorms and tornadoes, so scientists can't say whether climate change has made these events more likely. Reliable attribution is also difficult or even impossible where long-term climate records are still lacking, such as in some African countries. And there might still be natural climate variability that is not fully visible in the relatively short record of direct climate observations. To trace very-long-term climate fluctuations — such as those caused by changes in atmospheric-pressure patterns or sea surface temperatures that cycle once every few decades — researchers must rely on low-resolution proxy data, such as from tree rings. That this variability doesn't always show up in direct observations does create some uncertainty in studies, particularly for research on drought attribution, says Erich Fischer, a climate scientist at the Swiss Federal Institute of Technology in Zurich.

At a meeting in Oxford in 2012, some critics questioned whether climate scientists could be confident about the conclusions of attribution studies, given the lack of observational data and weaknesses in the climate models of the time. But since then, doubts have largely been quelled. Researchers now run the studies using several independent climate models, which reduces uncertainty because they can look for results that concur. And scientists are more careful about how they make probabilistic claims. "Extreme-event attribution has made a lot of progress since it began with scant resources," says Fischer. "It may still not work for small hailstorms or tornadoes. But attribution claims are now fairly robust for any large-scale weather patterns that can be represented by state-of-the-art climate models."

UNCLEAR IMPACT

In South Africa, Davies says Otto's latest study should help to press the case for new approaches to regional water management. "Meteorologists assured us after the second year of drought that there was no way we were going to have a third dry year in a row. But we can't use the past any more for what might happen in the future. We need to learn to adapt to a changing climate, and we absolutely need attribution to do it right." One of the lessons of the recent drought and the attribution analysis is that the Western Cape should not rely solely on rainfall to replenish its water supply, she says. Instead, it should diversify by tapping groundwater and expanding its desalination and waste-water treatment facilities.

But, in general, it's hard to know what effect attribution studies are having, social scientists say. That's because it is difficult to tease out the impacts of these findings from other studies that forecast increased risks of extreme weather associated with climate change — or from the shock of the weather events themselves. Still, if attribution studies start appearing regularly in weather reports, rather than just in scientific journals, then their impacts could become much more conspicuous, says Jörn Birkmann, an expert in spatial and regional planning at the University of Stuttgart in Germany. "City and infrastructure planners who plan and approve new housing areas, hospitals or train stations need to consider risks of extreme weather events more precisely if these events are clearly attributed to climate change," he says.

Evidence from attribution reports could also feed into litigation on climate change, suggest Birkmann and James Thornton, the London-based chief executive of ClientEarth, an international group of environmental lawyers. Court cases that allege failure to prepare for the effects of climate change haven't yet cited attribution studies, Thornton says. But he thinks judges will increasingly rely on them to help decide whether defendants — who might be oil companies, architects or government agencies — can be held liable. "Courts tend to give credibility to government data," he says. "If attribution moves from science to public service, judges will be much more comfortable using the results."

At the German weather agency, Becker says he is convinced that attribution studies will become a valuable service for many parts of society. "It's part of our mission to illuminate the links between climate and weather," he says. "There is demand for that information, there is science to provide it, and we are happy to spread it." ■

Quirin Schiermeier *is a senior reporter for* Nature *based in Munich, Germany.*

1. Stott, P. A., Stone, D. A. & Allen, M. R. *Nature* **432**, 610–614 (2004).
2. Otto, F. E. L. *Nature Clim. Change* **6**, 342–343 (2016).
3. Ray, A., Hughes, L., Konisky, D. M. & Kaylor, C. *Glob. Environ. Change* **46**, 104–113 (2017).
4. Imada, Y. *et al. Bull. Am. Meteorol. Soc.* **99** (Suppl.), S97–S101 (2018).
5. Knutson, T. R., Kam, J., Zeng, F. & Wittenberg, A. T. *Bull. Am. Meteorol. Soc.* **99** (Suppl.), S11–S15 (2018).
6. Walsh, J. E. *et al. Bull. Am. Meteorol. Soc.* **99** (Suppl.), S39–S43 (2018).
7. Dole, R. *et al. Geophys. Res. Lett.* **38**, L06702 (2011).
8. Rahmstorf, S. & Coumou, D. *Proc. Natl Acad. Sci. USA* **108**, 17905–17909 (2011).
9. Otto, F. E. L., Massey, N., van Oldenborgh, G. J., Jones, R. G. & Allen, M. R. *Geophys. Res. Lett.* **39**, L04702 (2012).

COMMENT

EVOLUTION The band of biologists who redrew the tree of life **p.26**



PHYSICS 'Whig' historian is back with reflections old and new **p.28**

SUSTAINABILITY Land-use change could threaten peace in Colombia **p.29**

HISTORY Children taught skewed view of science's past — fix it **p.29**

WENJIE DONG/GETTY



Late-night gridlock on Beijing's roads.

Set road charges in real time to ease traffic

Track vehicles to link tolls with demand and cut congestion, urge **Peter Cramton**, **R. Richard Geddes** and **Axel Ockenfels**.

Traffic jams cost us time, money and health. In 2016, the average US driver spent 42 hours in congestion during peak hours, and those in Los Angeles, California, spent 104 hours[1]. New Yorkers can walk almost as fast as vehicles crawl along streets in central Manhattan[2]. Being stuck is frustrating and stressful. Jammed cars can burn up to 80% more fuel than those in free traffic[3–5]. This leads to more air pollution and greater carbon dioxide emissions, increases the incidence of heart attacks, strokes and asthma and contributes to poor infant health[6,7], especially among city dwellers.

The economic damage of congestion last year in the United States, Germany and Britain totalled US$461 billion[8]. Such costs are rising as the world's population grows and urbanizes. The six most congested countries in 2017 were Thailand, Indonesia, Colombia and Venezuela, with Russia and the United States tied in fifth place.

The usual response is to call for more roads. But they don't diminish traffic[9]. Quite the opposite: more drivers move in. Nor will artificial-intelligence systems, ride-hailing services and autonomous cars ease the gridlock. Navigation systems draw more drivers to certain routes and can spread congestion to formerly quiet streets. The ride-hailing apps Uber and Lyft, for instance, have increased traffic because people make more car journeys. Without ride-hailing, according to one US survey, around half of such trips between 2014 and 2016 would have either not been made, or would have been done on foot, bicycle or public transport[10]. Similarly, self-driving cars use roads and fuel efficiently (and reduce accidents), but those gains might be ▶

© 2018 Springer Nature Limited. All rights reserved.



MUNSHI AHMED/BLOOMBERG/GETTY

Singapore has experimented with electronic pricing on roads in its inner zones.

swamped by an increased desire for cheap and easy transport[11].

The answer lies in dynamic road pricing. The location of individual cars can now be tracked to within a few centimetres. This makes it feasible to measure and price road use in real time according to demand. If the price were set at the right level, enough car drivers would choose to drive at a different time or take a different route or mode of transport to cut congestion. Limited road space would be managed in a similar way to airfares, electricity, hotel rooms and train journeys. Uber already balances demand and supply of its cars through surge pricing.

Overall, dynamic pricing does not drive motorists away. It can double the capacity of a congested route in peak times by preventing traffic jams[12] — just as managing fisheries can ease overfishing. Pollution and stress would decrease. The funds raised could be used to improve roads and public transport, and to reduce fuel and other taxes.

Fixed pricing schemes have been tried. Since 2015, 5,000 volunteers in Oregon have been trialling a tax on miles travelled by car. Around a dozen countries, including Germany, Austria and Switzerland, follow a similar approach for lorries. In recent decades, some cities in the United States and elsewhere, including Singapore and Stockholm, have experimented with electronic charging for roads in their inner zones.

But such schemes do little for congestion, because prices often do not change meaningfully with supply and demand[13]. A low price does little to mitigate jams at peak times. A price that is fixed high to eliminate peak congestion would be as inefficient and unacceptable as having Thanksgiving airfares all year.

Instead, policymakers and city managers need to track cars' positions and adjust charges continuously, depending on how busy the roads are.

What has stopped them? There are three research gaps: technology needs refining to measure road use by cars accurately and at low cost; an equity and privacy framework needs hammering out; and the implementation of network-wide, real-time road pricing needs better economic and computational modelling.

FREE RIDE

Congestion is pervasive because motorists take no account of the cost that they impose on others. Prices should instead reflect motorists' impact on each other. The system would operate in a similar way to an electricity market, making road space a commodity that can be bought and sold. An independent operator for the system would determine prices on each road segment to balance supply and demand, and thereby maximize the network's value to users while keeping traffic flowing[14].

Prices would be levied on all roads in a region. Charges could vary with time and place every ten minutes, say, according to traffic conditions. Prices would thus respond to lane closures, weather and sporting events, as well as to peak commuting times.

Research is needed to estimate the best levels for market prices. Most of the time, they would be near zero. On popular routes in Europe or the United States, an urban commute might cost as much as $20 — but the trip would take 30 minutes rather than, say, 45–90 minutes. The actual price would depend on how easily drivers can shift away to other times or modes of transport.

Fees would be tailored to vehicle types. Lorries would pay more. An autonomous vehicle, driven using algorithms to promote free flow and possibly coupled with other cars to decrease spacing, uses less road capacity than a standard car and would pay less.

An advantage of dynamic pricing is that it includes the means with which to charge the full social cost of a vehicle's use — both congestion and pollution. Prices could be varied to keep air-quality measures, such as particulate matter, within limits. Although this would not affect the number of cars at peak times, it would increase the road price for dirty vehicles and make room for clean ones. Electric cars, for instance, would pay less. This would be a cheaper and less-intrusive way of fighting pollution than banning diesel cars from city centres, as is being discussed and implemented in Germany.

Prices would be tracked using navigation apps such as Google Maps and Waze. Such tools would present both real-time

© 2018 Springer Nature Limited. All rights reserved.

information and forecasts of future prices, much as they do today for trip duration. Prices would be integrated into fares for taxis and ride-hailing services.

As in other markets, changes in pricing should be smooth to avoid price shocks. And consumers need time to react. Advance purchases, such as the ability to buy a pass for the daily commute, would let consumers plan and avoid the risk of expensive real-time prices. The operator for such purchases could offer packages as part of a road-use plan.

Not everyone will need to respond to prices. Dynamic pricing would still reduce congestion even if most drivers do not react. Yet more motorists would adapt as they became familiar with the system. Some would commute earlier or later; others would cycle or take the train or bus. In the long run, other adjustments, such as moving house, more work flexibility, investments in public transport and smarter mobility technologies, are possible. People already make similar trade-offs — the price of current road use is the length of delay.

There is much to learn about how best to manage a road network and how to target investments. Aggregate data about pricing and transport choices will need to be made publicly available. This would allow researchers and innovators to glean insights about the effectiveness of different measures and to develop apps that help motorists.

CONCERNS

Equity. Road-pricing schemes are often criticized as perpetuating inequality. People with lower incomes might be less able to afford to drive at popular times of day than those in wealthier groups. More research on such consequences is needed. But the problem could be smaller than feared. Pricing can make everybody better off — even before the revenue from congestion pricing is redistributed. For instance, suppose the left lanes of a busy multi-lane highway are priced at peak times. Because this increases throughput on the left lanes, there are fewer motorists on the right lanes, so everyone wins.

Even a worker with less money, who absolutely must be at work at 8 a.m. and who has no access to public transport or other travel options, can be better off — even if all lanes on all roads are priced. For example, suppose the free-flow travel time is 30 minutes, the expectation is 60 minutes and the maximum is 90 minutes. Today, without road pricing, the worker must depart at 6:30 a.m. to get there on time. With efficient pricing, if the worker cannot afford the price to depart at 7:30 a.m., she or he can continue to depart at 6:30 a.m. at zero cost, say, but his or her travel time will be halved and so fuel costs and pollution will be reduced. Moreover, the revenue from congestion pricing could be given back to motorists, for example, through a lower road tax and less fuel duty. It could also be invested to improve public transport.

The current situation, by contrast, is unfair. The free use of roads is equivalent to governments subsidizing people who impose the biggest congestion and pollution costs on society. Roads are an essential service. The norm for other essential services such as electricity, gas, water and communications is for consumers to pay for what they use. The fact that road use has not been priced is a fluke of history — until recently, technology did not allow for the measurement and communication of pricing.

Scepticism. Voters and politicians underestimate the benefits of pricing. Public support builds once people experience such schemes. For instance, before Stockholm introduced a €2 (US$2.3) charge during peak hours for vehicles entering the inner city, two-thirds of residents were against the plan. About 2 years later, when the policy had reduced traffic by 20%, two-thirds of people were in favour of it[15]. After a similar scheme was trialled in Milan, Italy, 80% of people voted to extend charges to more roads and vehicle types. More research is needed on how to communicate the impact of road pricing.

"The fact that road use has not been priced is a fluke of history."

Privacy. Road pricing raises concerns about personal data. Monitoring and enforcement require that the system operator know the location of each vehicle. Technically, this is easy and cheap: each vehicle would have a Global Positioning System (GPS) device. Strict privacy rules governing the operator imposing the charges would ensure that no individual data would be shared with others, as is standard in telecommunications and related industries. Modern cryptography makes it possible for the operator to run the market without any human having access to the data, yet prove that the rules are followed faithfully. Research is needed on whether such advanced systems give motorists comfort, although their embrace of apps for navigation and ride-hailing suggests that users are happy to accept some loss of privacy for improved services.

MAKE IT HAPPEN

The first step is to get devices in vehicles to measure road use. Singapore plans to install tracking and payment technology in all cars from 2020. Oregon's volunteers have GPS equipment on board so the state can levy a fee per mile. These are excellent starting points.

Governments and city authorities should establish independent operators to introduce and adjust prices for times and locations.

Researchers should study the impacts on motorists' behaviour, traffic conditions and pollution, the distribution of benefits and costs and public responses. This will involve studies by behavioural scientists of individual perceptions and changes in driving behaviour; by transport engineers, economists and computer scientists of aggregate traffic flows through the network; and by market designers, to evaluate the effectiveness of the underlying incentive mechanisms.

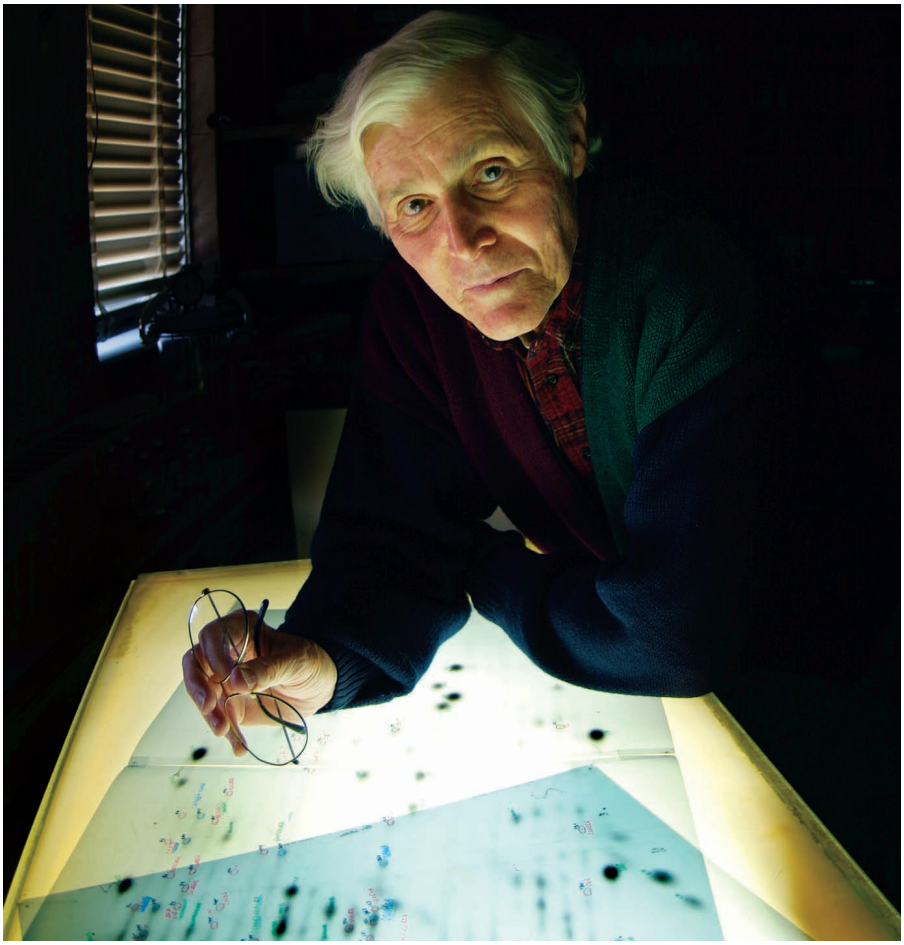
Service providers, such as Google, Apple, Uber and Lyft, are likely to seize the opportunity to develop innovative tools that enable consumers to make informed decisions. These tools would integrate past driving behaviour and relevant price information.

Let's get a move on. Dynamic pricing is the only way forward for roads. ■

Peter Cramton *is professor in the Department of Economics, University of Cologne, Germany.* **R. Richard Geddes** *is professor in the Department of Policy Analysis and Management, Cornell University, Ithaca, New York, USA.* **Axel Ockenfels** *is professor in the Department of Economics, University of Cologne.*
e-mail: ockenfels@uni-koeln.de

1. Pishue, B. *US Traffic Hotspots: Measuring the Impact of Congestion in the United States* (INRIX Research, 2017).
2. Berger, P. & Vilensky, M. 'Push for New York Congestion Charge Picks Up Steam' *The Wall Street Journal* (1 January 2018).
3. Borut, J., Samo, K., & Tadej, B. *FME Trans.* **46**, 278–284 (2018).
4. Treiber, M., Kesting, A. & Thiemann, C. *How Much Does Traffic Congestion Increase Fuel Consumption and Emissions?* Annual Meeting of the Transportation Research Board, Washington DC (2008); available at go.nature.com/2urgk9f
5. Garcia-Castro, A. & Monzon, A. in *Transportation Research Board 94th Annual Meeting Compendium of Papers* (TRB, 2015); available at go.nature.com/2zwbzkt
6. Currie, J. & Walker, R. *Am. Econ. J. Appl. Econ.* **3**, 65–90 (2011).
7. Levy, J. I., Buonocore, J. J. & von Stackelberg, K. *Environ. Health* **9**, 65 (2010).
8. 'The Hidden Cost of Congestion' *The Economist* (28 February 2018).
9. Duranton, G. & Turner, M. A. *Am. Econ. Rev.* **101**, 2616–2652 (2011).
10. Clewlow, R. R. & Mishra, G. S. *Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States* (Institute of Transportation Studies, 2017); available at go.nature.com/2ugicvh
11. Fulton, L., Mason, J. & Meroux, D. *Three Revolutions in Urban Transportation* (UC Davis Institute of Transportation Studies, 2017); available at go.nature.com/2upbcxc
12. Federal Highway Administration. *Congestion Pricing — A Primer: Overview* (US Department of Transportation, 2008); available at go.nature.com/2jkku2v
13. Martin, L. A. & Thornton, S. *Can Road Charges Alleviate Congestion?* Working Paper (Univ. Melbourne, 2017); available at go.nature.com/2ma65ep
14. Cramton, P., Geddes, R. R. & Ockenfels, A. *Markets for Road Use: Eliminating Congestion through Scheduling, Routing, and Real-time Road Pricing* Working Paper (Univ. Cologne, 2018); available at go.nature.com/2zwe2qc
15. Eliasson, J. *The Stockholm Congestion Charges: An Overview* Working Paper 2014, 7 (Centre for Transport Studies Stockholm, 2014); available at go.nature.com/2obp4mg

© 2018 Springer Nature Limited. All rights reserved.



JASON LINDLEY/LAS, UNIV. ILLINOIS AT URBANA-CHAMPAIGN

Carl Woese discovered the 'third domain' of life — the Archaea.

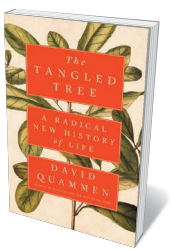
EVOLUTION

Scaling life's tree

John Archibald praises a compelling guide to 3 billion years of life — and its molecular historians.

In *The Tangled Tree*, celebrated science writer David Quammen tells perhaps the grandest tale in biology: how scientists used gene sequencing to elucidate the evolutionary relationships between living beings. Charles Darwin called it the 'great Tree of Life'. But as Quammen reveals, at the molecular level, life's history is more accurately depicted as a network, a tangled web through which organisms have been exchanging genes for more than 3 billion years. This perspective is indeed radical, and he presents the science — and the scientists involved — with patience, candour and flair.

Centre stage in Quammen's narrative is Carl Woese (1928–2012), the US microbiologist best known as the discoverer of the Archaea (Archaebacteria) — the 'third domain' of life. Inspired by the visionary musings of Francis Crick, Linus Pauling and Emile Zuckerkandl, Woese committed himself to molecular phylogenetics at a time when this powerful approach to the study of evolution was in its infancy. During the 1960s and 1970s, the Woese Laboratory at the University of Illinois at Urbana–Champaign developed and refined techniques for deriving sequence information from molecules of ribosomal RNA (core components of the cell's protein-synthesizing factory, the ribosome). Sequences were painstakingly obtained from diverse microbes and used as molecular yardsticks to infer how the organisms were related to one another and to animals and plants. Through the following two decades, as molecular sequencing got easier and cheaper, Woese's 'three-domains' tree — comprising archaea, bacteria and the nucleus-containing eukaryotes — served as the definitive road map for the field of comparative genomics. In many ways, it still does.



The Tangled Tree: A Radical New History of Life
DAVID QUAMMEN
Simon & Schuster (2018)

But life is complicated, and so are the scientists who study it. In his breezy, conversational style, Quammen shepherds us up and down life's vast timeline, and across 150-plus years of exciting, often controversial discoveries. He handles the complexities with humour and clarity (he's right: some ribosomes do look like rubber ducks). We learn about the seeds of "tree thinking" in biology, before and after Darwin's 1859 *On the Origin of Species*. We learn of a time when a natural classification of microorganisms was considered impossible (they were deemed morphologically too simple, physiologically too variable). We learn how molecular sequencing helped test and eventually prove the endosymbiont hypothesis for the origin of mitochondria and chloroplasts; these eukaryotic organelles are now known to have evolved from once free-living bacteria.

And we learn that although molecular phylogenetics provided the means with which to build a universal tree of life that includes microbes, it also provided the data that ultimately led us to question the precise nature of the tree. From the late 1990s onwards, with dozens and eventually thousands of complete genome sequences in hand, biologists began to realize that the horizontal exchange of genes between distantly related organisms is an important evolutionary force. (Quammen also reminds us that, as early as 1963, medical microbiologist Tsutomu Watanabe and colleagues provided evidence for horizontal gene transfer as a mediator of antibiotic resistance in bacteria.) Because genes have "moved sideways", not all genes in a given genome share the same history. Current evidence suggests that this is also true for at least some macroorganisms (such as plants). The tree of life is tangled, some branches hopelessly so.

At times, this master storyteller's book reads like a travelogue. It brims with revelations from dozens of interviews with key players in their native habitats: the late Lynn Margulis, US champion of endosymbiotic theory; former Woese Lab members George Fox, Mitchell Sogin and Linda Bonen; and the environmental-DNA-sequencing legend and three-domains defender Norman Pace. Here too are "the four horsemen" of the gene-transfer apocalypse: William Martin, Jeffrey Lawrence, Peter Gogarten and Ford Doolittle.

Some of the stories are laugh-out-loud

© 2018 Springer Nature Limited. All rights reserved.

funny. In one, Woese's collaborator Charles Vossbrinck — an "openhearted bear of a man" — picked up a tipsy, pontificating Woese at a barbecue and threw him into the bushes. (Their friendship survived.) Other tales are shockingly intimate. Woese's last months and weeks with pancreatic cancer, as revealed by those closest to him, make for painful, albeit illuminating reading. I was surprised, for instance, to learn that Woese believed in a deity.

The Tangled Tree traces the full arc of Woese's life and career. We see the fiercely determined young scientist struggling to collect the data that he intuited would be important, and the brooding, combative mid-career professor fighting to have his beloved archaea and three-domains tree accepted by the scientific community. Finally, there is the jaded, curmudgeonly legend wracked by a Darwin complex. None of the accolades showered on Woese seemed to matter (he and many others clearly felt he deserved a Nobel prize, but he never got one). Around 2010, Woese and Canadian science historian Jan Sapp began to collaborate on a book tentatively entitled *Beyond God and Darwin*. The project never moved beyond Sapp's draft introduction, on which Woese wrote: "Jan, you accord Darwin so much more substance than the bastard deserves."

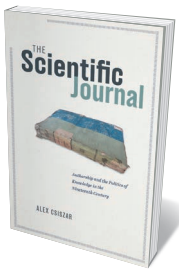
"Life is complicated, and so are the scientists who study it."

Above all, Quammen reminds us that science is an imperfect, highly social activity. It happens in labs — but also in hallways and airports, over pizza or coffee. And as with any other human endeavour, egos and reputations play a huge part. Friendships are forged, broken and mended over perceived or actual slights in the literature or at conferences. The actual data matter less often than we would like to admit.

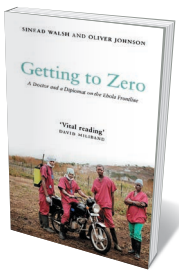
To what extent is the tree metaphor still 'useful'? On this thorny question, Quammen is clear: among practising scientists, opinions differ greatly. Horizontal gene transfer is here to stay — it's now a question of how, how much, how important and between which organisms. And it is here that our twenty-first-century science connects back to the centuries-old struggle to classify and make sense of the world around us. At root, science and philosophy are interwoven in ways that many of us fail to realize, a fact to which Quammen is wisely alert. ■

John Archibald *is a professor in the Department of Biochemistry & Molecular Biology at Dalhousie University in Halifax, Canada, and the author of* One Plus One Equals One: Symbiosis and the Evolution of Complex Life.
e-mail: jmarchib@dal.ca

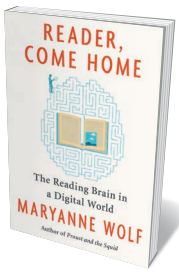
Books in brief



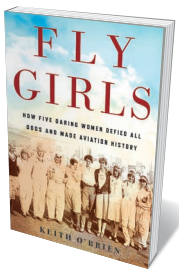
The Scientific Journal
Alex Csiszar Univ. Chicago Press *(2018)*
Journals form the canon of scientific knowledge. But how, asks historian Alex Csiszar, did it come to bear "so much epistemic weight"? Focusing mainly on France and Britain during the turbulent nineteenth century, he unpicks the knotted roots of journals from the Royal Society's *Philosophical Transactions* to *Annales des sciences naturelles*, and touches on the role of luminaries such as *Nature*'s first editor, Norman Lockyer. Amid fresh convulsions in scholarly publishing, much here resonates — not least, how commercial interests have shaped science communication almost from the start.



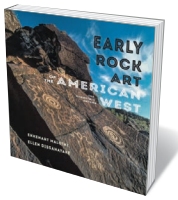
Getting to Zero
Sinéad Walsh and Oliver Johnson Zed *(2018)*
In 2014, as West Africa's Ebola crisis exploded, 28-year-old physician Oliver Johnson was co-running the isolation unit in Sierra Leone's main hospital; Sinéad Walsh was Irish ambassador to the country and head of Irish Aid. Their in-depth memoir enshrines distinct perspectives on the front line of a fraught epidemic, to offer a nuanced analysis: we see both the Herculean efforts on the ground, and the humanitarian response, warts and all (see also P. Piot *Nature* **537**, 484–485; 2016). Among the lessons learnt, the need to respect local 'citizen medics' and collaborate with governments is pure gold.



Reader, Come Home
Maryanne Wolf Harper *(2018)*
This rich study by cognitive scientist Maryanne Wolf tackles an urgent question: how do digital devices affect the reading brain? Wolf explores the "cognitive strata below the surface of words", the demotivation of children saturated in on-screen stimulation, and the power of 'deep reading' and challenging texts in building nous and ethical responses such as empathy. She advocates "biliteracy" — teaching children first to read physical books (reinforcing the brain's reading circuit through concrete experience), then to code and use screens effectively. An antidote for today's critical-thinking deficit.

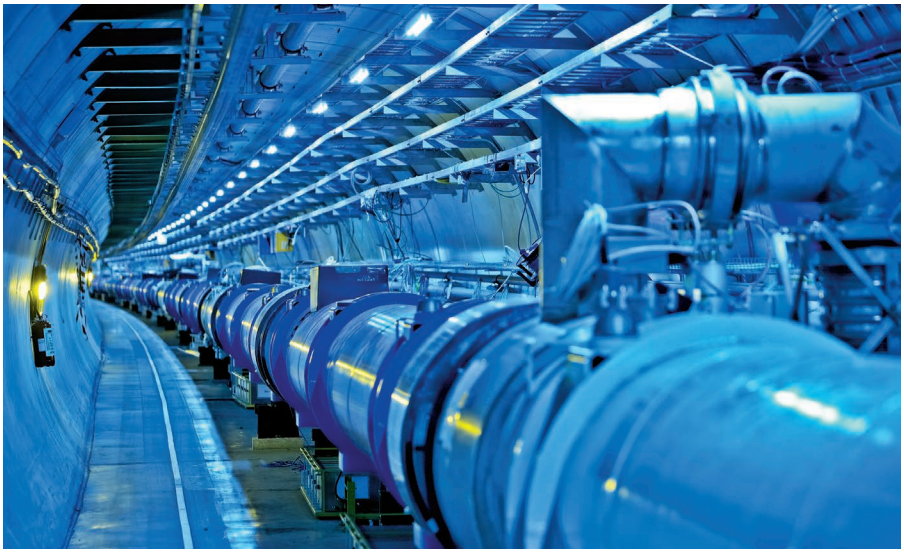


Fly Girls
Keith O'Brien Houghton Mifflin Harcourt *(2018)*
Shredded wings, broken propellers, stalled engines: in the 1920s, aviation was insanely risky. Undeterred, a select cadre of women embraced US aeroplane racing. In this engrossing mix of group biography and technology history, Keith O'Brien follows the lives of five: Amelia Earhart, Florence Klingensmith, Ruth Elder, Ruth Nichols and Louise Thaden. Earhart became a celebrity before disappearing over the Pacific Ocean; others found their prowess no match for sexism. The brilliant record-breaker Nichols, for instance, never flew professionally after the Second World War, and killed herself in 1960.



Early Rock Art of the American West
Ekkehart Malotki and Ellen Dissanayake Univ. Washington Press *(2018)*
The ancient geometric petroglyphs and pictographs of the American West — pecked into or painted on boulders and canyon walls — are beautiful enigmas. In this fascinating volume, linguist Ekkehart Malotki and scholar Ellen Dissanayake parse images created up to 15,000 years ago by Palaeoamericans from Arizona to Idaho, speculating about their origins and functions. Alongside Malotki's stunning photographs of some 200 examples, the authors recontextualize the relics as products of ritualistic activity ('artification') rather than symbolic artworks. **Barbara Kiser**

© 2018 Springer Nature Limited. All rights reserved.



THOMAS PFLAUM/VISUM/EYEVINE

Part of the ring tunnel at the Large Hadron Collider near Geneva, Switzerland.

PHYSICS

A rational reductionist argues his case

Robert Crease enjoys an eminent physicist's turn as the ideal witness to science and its history.

Third Thoughts
STEVEN WEINBERG
Belknap (2018)

In his 1974 book *The First Three Minutes* (1977), physicist and Nobel laureate Steven Weinberg pictured his ideal reader as "a smart old attorney" who might not know much science, but "expects nonetheless to hear some convincing arguments". That's still his approach in *Third Thoughts*. This essay collection (his third for a lay readership, hence the title), ranges widely over science, the history of science and current affairs, taking on everything from dark energy and quantum field theory to socio-economic inequity and the wastefulness of crewed space missions.

The volume mixes new pieces with others previously published in *The New York Review of Books*. Weinberg acknowledges the help he received from the review's formidable editor, the late Robert Silvers. He clearly apprenticed well: if his ideal reader is an astute lawyer, Weinberg might be described as an ideal witness. He is clear, to the point, frank and transparent about his perspective — "rationalist, realist, reductionist, and devoutly secular".

Among half a dozen pieces on particle physics are lucid explications of, for instance, the Higgs boson, Hilbert space and the Large Hadron Collider. Weinberg has a knack for capturing a complex concept in a succinct, unforgettable image. He compares quantum superposition, in which a particle has two states at once, to a musical chord; when measurement collapses the particle to one state, it "somehow shifts all the intensity of the chord to one of the notes, which we then hear on its own". And he describes the discovery that nature obeys symmetries whose consequences can be worked out as "like having a spy in the enemy's high command".

A few of the essays delve into history. Weinberg has irked professionals in the field by venturing into their territory, but not because he gets his facts wrong. As a self-confessed 'Whig' historian, he believes that the past should be judged by values of the present (the term springs from the name of the long-defunct British political party whose members thought history had been building towards Parliamentary government). Weinberg offers a curious panorama that demotes certain canonical figures, such as Democritus, Francis Bacon and René Descartes.

The ancient Greek philosopher Democritus didn't make observations, so, although he correctly proposed that matter is made of atoms, Weinberg argues that he "was wrong about how to learn about the world". Yet Democritus is of supreme importance for the understanding of science history (as Weinberg admits) because he inspired atomic theorists of the early modern period, such as Robert Boyle and other corpuscular philosophers, in their efforts to explain nature without using theology or teleology. Similarly, Bacon and Descartes, although sometimes scientifically misguided by today's standards, were vastly important for ushering in the mechanical thinking that Weinberg himself practises. Two essays here convey Weinberg's responses to critics of his Whig convictions, but I wish he had put the arguments and counter-arguments more forcefully and at greater length.

Weinberg is at his most interesting when probing the big uncertainties in physics. Ultimately, he is not sure, for instance, that he knows what an 'elementary particle' actually is, or how best to interpret quantum mechanics. These moments reveal Weinberg's considerable integrity, where — as one of the smartest and most diligent scientists around — he describes himself as somewhat lost. He has surveyed the present, and all the best paths forward proposed by other scientists; yet, at gut level, he is confident that nothing on the horizon is fully satisfactory, and that other possibilities might be out there. These admissions imply that Weinberg suspects future historians may harbour a perception of today's thinking very different from ours.

The articles on public and personal matters — Weinberg's thoughts about taxes, his disappointment with former US President Barack Obama for failing to confront economic disparities more directly, and educators who are honoured with burial in Texas State Cemetery — are less interesting. Yet here, as elsewhere, he is clever: "The only technology for which the manned space flight program is well suited is the technology of keeping people alive in space. And the only demand for that technology is in the manned space flight program itself."

I read the penultimate essay with anticipation. Weinberg reveals in the introduction that he had not published it before because nobody who read it liked it. You can see why. It's an earnest piece of amateur philosophizing that compares creativity in theoretical physics to that in the arts, building on Weinberg's feeling that success in each depends on "a sense of inevitability". He would have been clearer had he consulted a few philosophical concepts, such as Immanuel Kant's idea that a beautiful work creates the feeling that it has the inevitability of a product of nature. But Weinberg (whose 1992 book *Dreams of a Final Theory* contains a chapter entitled 'Against Philosophy') was not about to take that route. In all these essays, his directness both enlightens and illuminates flaws in his own arguments. Witnesses, of course — even ideal ones — have blind spots. The best thing about Weinberg's essays is that they do, indeed, make you feel like a smart old attorney. ■

Robert P. Crease *is a professor in and chair of the Department of Philosophy at Stony Brook University, New York.*
e-mail: robert.crease@stonybrook.edu

© 2018 Springer Nature Limited. All rights reserved.

Correspondence

Make databases language-proof

It is absurd to put effort and public resources into research that has already been published. This will continue to be a risk as long as papers in non-English journals are not routinely indexed in the international databases (see also J. Lebel and R. McLean *Nature* **559**, 23–26; 2018).

Some national databases offer a partial solution (see J. Tao *et al. Nature* **557**, 492; 2018). For example, Ukraine's Panteleimon database (http://www.panteleimon.org) translates the title, abstract and some figure legends and tables into English. Nevertheless, people should never cite research that has not been read in full.

The scientific community needs to develop a comprehensive multi-language translation tool with the help of services such as Google Translate. This would enable international researchers to access regional databases not compiled in English and to find out all the essential details — for instance, regarding experimental design and results, or whether the paper was peer-reviewed. It would also resolve problems of priority and giving proper credit.

Daniel Prieto *Instituto de Investigaciones Biologicas Clemente Estable, Montevideo, Uruguay.*
dprieto@fcien.edu.uy

Beyond replicability in the humanities

The humanities should take responsibility for quality in the same way the sciences do, argue Rik Peels and Lex Bouter, through the pursuit and institutionalization of replicability (*Nature* **558**, 372; 2018). We disagree: quality criteria are crucially different in the humanities and the sciences.

The humanities pursue meaning beyond truth. Confirming that Van Gogh painted *Sunset at Montmajour* (truth) is only the beginning. Unearthing the cultural meaning of the work requires historical context and theorizing on its message, style, aesthetics — and what the work can tell us about the artist and his world (view). The coexistence of multiple valid answers and the value of their interaction disqualify replication as a viable quality criterion.

Moreover, the humanities relate differently to their objects of study. They focus on both interactive kinds (people) and indifferent kinds (atoms, DNA sequences, paintings). Extracting meaning from interactive data requires continued interaction between informants, who might resist or embrace preliminary results or classifications. With co-producers of data and meaning, protocols are never set in stone, reporting guidelines are necessarily local and consent is always fluid.

Replication is a mark of quality only in the construction of truth for indifferent kinds. Extracting meaning from interactive kinds requires evaluation and assessment according to different quality criteria — namely, those that are based on cultural relationships and not statistical realities.

Sarah de Rijcke *Leiden University, Leiden, the Netherlands.*
Bart Penders *Maastricht University, Maastricht, the Netherlands.*
s.de.rijcke@cwts.leidenuniv.nl

Help relieve poverty with solar power

Of China's ten poverty-alleviation projects, its development of photovoltaic-based solar power has been one of the most successful. We suggest that other countries look more explicitly at solar energy as a way of generating income in rural areas, in accord with the United Nations Sustainable Development Goal to eradicate global poverty by 2030.

China's overall programme has lifted more than 50 million rural people out of poverty since 2013 (Y. Zhou *et al. Land Use Policy* **74**, 53–65; 2018). Solar-energy schemes launched in 2014 supplied 7.9 gigawatts of power by the end of 2017, directly benefiting some 800,000 poverty-stricken families (see go.nature.com/2jtdxjh; in Chinese). In Lixin county in central China, for example, solar installations provided an additional annual income of more than 3,000 yuan (around US$440) for every family.

Solar-power facilities provide employment opportunities, boost farmers' incomes and supply households with affordable, reliable and sustainable energy, thus also helping to alleviate energy poverty.

Yang Zhou, **Yansui Liu** *Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China.*
liuys@igsnrr.ac.cn

Land use must abide by peace agreement

A resolution signed in June to allow agricultural development on 35% (40 million hectares) of Colombia's land could risk compromising the government's 2016 Peace Agreement with the Revolutionary Armed Forces (see also *Nature* **558**, 169–170; 2018). The agreement places strict controls on the transformation of national lands and environmentally important areas.

At present, just 20% of that land is under cultivation. How the other 80% may be used is unspecified, but we fear that ecologically friendly farming and traditional production systems — such as cattle ranching in flooded savannahs in the Orinoquía region — are likely to be replaced by more-intensive forms of land exploitation.

The expansion threatens the peace process and prospects for sustainable rural development — already a challenge in a country where only 16% of the soil is legally protected against degradation (see go.nature.com/2v997uv).

Luca Eufemia, **Michelle Bonatti** *Leibniz-Centre for Agricultural Landscape Research, Müncheberg, Germany.*
Marcos A. Lana *Swedish University of Agricultural Sciences, Uppsala, Sweden.*
luca.eufemia@zalf.de

Rectify biased take on science history

As members of the STEM Advocacy Institute, we find the typical Western view of science history distorted and incomplete and argue for more-balanced representation. Many non-Western scientists have made hugely important contributions to scientific knowledge, but their rich and inspiring stories garner little attention in the West.

For example, Hippocrates is widely considered to be the 'father of medicine' — even though the ancient Egyptians developed medicine as a profession 2,000 years earlier (see www.ancient.eu/imhotep). The first known physician in Egypt was a man named Imhotep, who was deified after his death for his medical achievements (see go.nature.com/2uxs5qd). Many such examples exist, but they are not well-known (see, for instance, J. Al-Khalili *Nature* **518**, 164–165; 2015; A. M. Celâl Şengör *Nature* **471**, 162–163; 2011; J. Poskett *Nature* **550**, 332; 2017).

This means that schoolchildren are inculcated with a history that excludes the diversity of ethnicities, beliefs and cultures that have contributed to today's science, technology, engineering and mathematics. Ignoring these reinforces stereotypes and the marginalization of certain groups, whereas balancing the narrative would positively influence those who are already disadvantaged in our classrooms.

Aiza Kabeer *Manchester, New Hampshire, USA.* **Jessica W. Tsai** *Boston, Massachusetts, USA.*
scholarship@stemadvocacy.org

© 2018 Springer Nature Limited. All rights reserved.

CARBON CYCLE

Microbes weaken soil carbon sink

The rate at which carbon dioxide is lost from soil has risen faster than the rate at which it is used by land plants, because soil microbes have become more active — possibly weakening the land surface's ability to act as a carbon sink. SEE LETTER P.80

KIONA OGLE

The terrestrial land surface has a crucial role in the global carbon cycle, providing feedbacks to changes in atmospheric levels of carbon dioxide and associated climate change[1]. Increases in atmospheric CO_2 concentrations and in soil and air temperatures worldwide over the past several decades have been paralleled by an increase in the metabolism of organisms at the land surface — as demonstrated by enhanced rates of CO_2 uptake, mainly by plants through photosynthesis, and of CO_2 loss from plants and soil microorganisms, mostly owing to respiratory processes[2–6]. On page 80, Bond-Lamberty *et al.*[7] report that the rate of increase of CO_2 loss is outpacing that of CO_2 uptake by plants. The authors attribute the imbalance in these rates of increase to enhanced activity of microbes that obtain nutrition by decomposing or mineralizing organic matter in soil. If the observed trend continues, then respiration by microbes could contribute substantially to global warming by releasing CO_2 from organic matter that has previously been stored in soil for decades to millennia.

A variety of processes underlie the exchange of CO_2 between the land surface and the atmosphere. Bond-Lamberty *et al.* focused on soil respiration, which is arguably one of the largest fluxes of CO_2. The authors analysed previously published soil-respiration data[8] from many sites around the world that covered a broad range of ecosystems, including cropland, temperate forest and desert. They used these data to estimate the annual rates of soil respiration at various sites and to evaluate trends between 1990 and 2014.

Bond-Lamberty and colleagues then compared trends in soil respiration (CO_2 loss) with those of plant productivity (CO_2 uptake) that were derived from different data sources, including satellites. They found that the ratio of the rate of soil respiration to that of plant productivity has, in general, increased over the period covered by their data set. The ratio rarely exceeded 1, except at certain sites in particular years, which indicates that specific situations can lead to more CO_2 being lost from soil than is taken up by plants.

The findings beg the question of whether the average global ratio could become greater than 1 in the future, and, if so, when? Such an event would mark the tipping point at which the land surface stops operating mainly as a sink that helps to remove atmospheric CO_2 that is derived from fossil-fuel emissions[9,10], and starts acting as a source of CO_2 — exacerbating rising CO_2 levels and accelerating the pace of climate change[11] (Fig. 1).

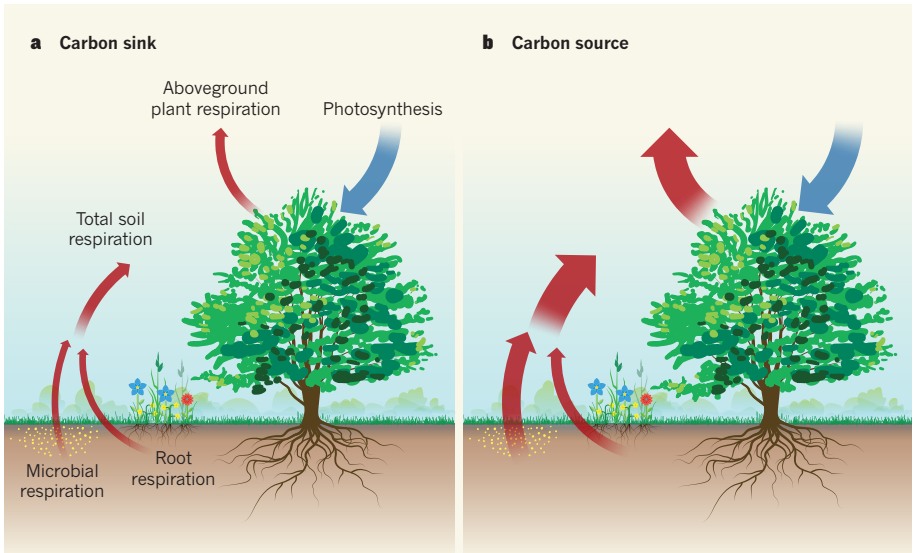


Figure 1 | Tipping the balance of carbon fluxes at Earth's land surface. **a**, The main sources of carbon dioxide from the terrestrial biosphere are respiration that is associated with the decomposition of organic matter in soil by microbes, and respiration of plants (both aboveground and belowground). Plants also absorb CO_2 through photosynthesis. At present, the total amount of CO_2 that is absorbed by plants exceeds the amount that is produced by respiration, and so the land surface acts as a carbon sink. Arrow widths roughly correspond to the sizes of the CO_2 fluxes. **b**, Bond-Lamberty *et al.*[7] report that, over the past few decades, the rate at which CO_2 is produced by soil microbes has increased faster than that at which CO_2 is used by plants. This raises the possibility that the rate of respiration will reach a tipping point at which it overtakes the rate of CO_2 uptake by plants. Under such a scenario, the land surface would act as a source of atmospheric CO_2. The time at which such a tipping point would be reached is unclear.

The authors next focused on studies in their data set that broke down total soil respiration into respiration dominated by decomposition by microbes and that associated with plant roots. Analysis of the microbial-dominated respiration rates led them to conclude that the disproportionately faster increase in the rate of total soil respiration is due to the enhanced activity of soil microbes. However, to understand whether accelerated rates of soil respiration will cause the land surface to become a source of CO_2, the temporal trends in respiratory losses associated with aboveground plant biomass must also be considered — the total loss of biologically derived CO_2 from the terrestrial biosphere is the sum of the soil and non-soil losses.

As Bond-Lamberty *et al.* acknowledge, previously published long-term data[12,13] recorded by eddy-covariance towers, which continuously monitor CO_2 concentrations and fluxes at specific sites across a range of ecosystems, suggest that the rate of increase of plant productivity has been faster than that of the total aboveground and belowground respiratory CO_2 losses. Further data and analyses are required to explain why those findings apparently contradict the authors' results.

If Bond-Lamberty and colleagues' findings are correct, which mechanisms could explain the markedly enhanced stimulation of the activity of soil microbes relative to plant productivity and plant respiration? Studies in the past few years have shown that the ability of plants to downregulate respiration in response to long-term increases in temperature[14] is

© 2018 Springer Nature Limited. All rights reserved.

much greater than that of short-lived soil microbes[4,15–17]. The authors suggest that the increased microbial activity observed in their study probably reflects the stimulatory effects of elevated temperatures associated with climate change.

There are, however, potential issues when drawing global inferences from the data analysed by Bond-Lamberty and co-workers. Most of the data came from spot measurements of soil-respiration rates that were obtained by many different researchers, who used a variety of methods to work out the contributions of soil microbes. This diversity of methods might have led those researchers to come to contrasting conclusions about the relative importance of soil microbes in their studies. Moreover, Bond-Lamberty *et al.* used simplifying assumptions to translate hourly or daily snapshots of respiration rates into annual fluxes of CO_2, but did not take into account the uncertainty in these calculations. The soil-respiration data set is also limited in its temporal coverage of individual sites: repeated observations were available for only a handful of sites, yet recurrent observations are necessary to prevent temporal trends from being obscured by factors that vary between sites.

The authors acknowledge and account for some of these limitations in their statistical analyses, but clearly there is room for a more rigorous investigation. This would require researchers to gather continuous time series of soil respiration and its component fluxes, and demands the use of precise methods for quantifying uncertainty and for extrapolating local measurements to determine trends in larger regions. Despite the limitations, Bond-Lamberty and colleagues' work is valuable because it aids our understanding of soil's long-term potential for sequestering carbon — as well as how this sequestration might be threatened by accelerated rates of organic-matter decomposition by soil microbes. Their findings will be crucial for developing and testing models of the global carbon budget, of which soil carbon is a central component.

Fluxes of CO_2 across whole ecosystems are often measured using eddy-covariance towers. By contrast, continuous measurements of soil respiration and decomposition by microbes are not broadly available for sites worldwide or do not cover multi-year periods. The establishment of long-term observational projects such as the US National Ecological Observatory Network (NEON), which monitors fluxes of soil CO_2 among other ecological measures, will create opportunities for the systematic evaluation of temporal trends and the underlying causes of changes in the rates at which CO_2 is lost from soil. Such data will be paramount for developing regional and global models of the carbon cycle, as well as for assessing climate change and the strategies by which it might be mitigated[18]. ■

Kiona Ogle *is in the School of Informatics, Computing, and Cyber Systems, and in the Department of Biology and the Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, Arizona 86011, USA.*
e-mail: kiona.ogle@nau.edu

1. Green, J. K. *et al. Nature Geosci.* **10**, 410–414 (2017).
2. Lu, M. *et al. Ecology* **94**, 726–738 (2013).
3. Piao, S. *et al. Nature* **451**, 49–52 (2008).
4. Wang, X. *et al. Glob. Change Biol.* **20**, 3229–3237 (2014).
5. Zhou, L. *et al. Glob. Change Biol.* **22**, 3157–3169 (2016).
6. Zhu, Z. *et al. Nature Clim. Change* **6**, 791–795 (2016).
7. Bond-Lamberty, B., Bailey, V. L., Chen, M., Gough, C. M. & Vargas, R. *Nature* **560**, 80–83 (2018).
8. Bond-Lamberty, B. & Thomson, A. *Nature* **464**, 579–582 (2010).
9. Le Quéré, C. *et al. Nature Geosci.* **2**, 831–836 (2009).
10. Pan, Y. *et al. Science* **333**, 988–993 (2011).
11. Wieder, W. R., Cleveland, C. C., Smith, W. K. & Todd-Brown, K. *Nature Geosci.* **8**, 441–444 (2015).
12. Keenan, T. F. *et al. Nature* **499**, 324–327 (2013).
13. Urbanski, S. *et al. J. Geophys. Res-Biogeo.* **112**, G02020 (2007).
14. Reich, P. B. *et al. Nature* **531**, 633–636 (2016).
15. Karhu, K. *et al. Nature* **513**, 81–84 (2014).
16. Oliverio, A. M., Bradford, M. A. & Fierer, N. *Glob. Change Biol.* **23**, 2117–2129 (2017).
17. Schindlbacher, A., Schnecker, J., Takriti, M., Borken, W. & Wanek, W. *Glob. Change Biol.* **21**, 4265–4277 (2015).
18. Walsh, B. *et al. Nature Commun.* **8**, 14856 (2017).

AUTOIMMUNITY

Peptide secretion triggers diabetes

An autoimmune attack on cells that make the hormone insulin causes type 1 diabetes. A mouse study reveals that pancreatic-β-cell release of insulin peptide fragments into the bloodstream triggers this harmful process. SEE LETTER P.107

JIAJIE WEI & JONATHAN W. YEWDELL

Diabetes arises from problems in the regulation of blood glucose, which is controlled by releasing the hormone insulin. The amount of insulin made is abnormally low in type 1 diabetes owing to the autoimmune-mediated destruction of insulin-producing β-cells in the pancreas. Wan *et al.*[1] reveal on page 107 how this immune attack is triggered.

Type 1 diabetes was a lethal condition with a life expectancy of just months until the discovery of insulin in 1921 enabled clinical management by insulin injection[2]. Although this therapy greatly extends life expectancy, a deeper understanding of the disease is needed to develop treatments that delay or prevent disease onset.

Studies of non-obese diabetic (NOD) mice, which spontaneously develop the disease, have provided insights into the mechanisms that cause the condition. Such work has revealed that T cells of the immune system have a key role in destroying insulin-producing pancreatic β-cells. T-cell immunosurveillance is aided by antigen-presenting cells, which present peptide fragments called antigens on their surface bound to major histocompatibility complex (MHC) class I and II molecules. T cells typically encounter antigen-presenting cells in the lymph nodes, and their T-cell receptor samples the antigens presented on MHC molecules. T cells that respond to self-antigens are normally eliminated as they mature in the thymus gland, but imperfections in this process can lead to autoimmunity. Although many proteins present in pancreatic β-cells could potentially provide the autoimmune trigger for type 1 diabetes, insulin is the culprit in NOD mice[2].

In type 1 diabetes in humans and NOD mice, the presentation of a peptide consisting of amino acids 12 to 20 of insulin's B chain (B:12–20) by MHC class II molecules can activate CD4 T cells that recognize this peptide[3]. In NOD mice, a class II MHC molecule called I-A^{g7} presents B:12–20 to CD4 T cells and activates them. These cells then initiate a process that activates CD8 T cells, which are specific for other β-cell peptides[4]. Activated CD8 cells kill β-cells, leading to diabetes when the T cells eliminate enough β-cells. The risk of a person developing type 1 diabetes is often linked to MHC class II genes, which are among the mostly highly variable human genes. The gene encoding the version of MHC class II called HLA-DQ8 is tightly associated with the disease[5]. Remarkably, HLA-DQ8 has a highly similar peptide-binding specificity to that of I-A^{g7} (ref. 6).

"Why isn't there selection against the generation of secreted insulin peptides that trigger autoimmunity in humans?"

Wan and colleagues confirmed the pathological potential of T cells that are specific for I-A^{g7}–B:12–20 complexes by introducing such T cells into NOD mice under conditions that generate enough CD4 T cells to cause type 1 diabetes. As a control, the authors did a similar transfer into NOD mice that have a

© 2018 Springer Nature Limited. All rights reserved.

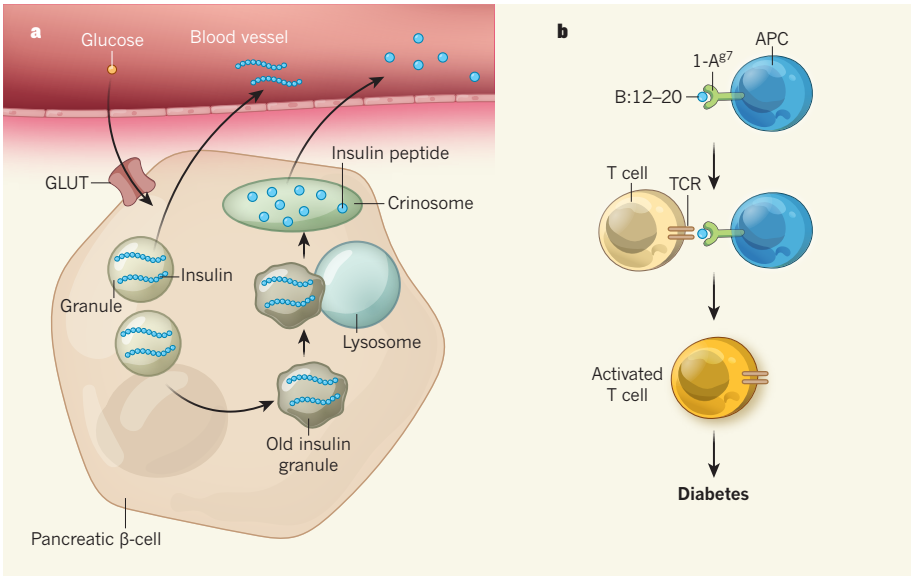


Figure 1 | Insulin and type 1 diabetes. If T cells launch an autoimmune attack on pancreatic β-cells, which produce insulin protein, it causes diabetes. The source of the specific insulin peptides that can prime the T cells that give rise to diabetes has remained unknown. **a**, Wan *et al.*[1] have now identified this source. Glucose from the blood enters β-cells through the GLUT channel protein, triggering the release of insulin from granules into the blood. The group found that old insulin granules are destroyed by fusing with an organelle called the lysosome to form another organelle known as the crinosome, in which insulin peptides arise. The peptides enter the bloodstream together with full-length insulin after glucose influx. **b**, The authors report that lymph nodes (not shown) are a site where an insulin peptide that might trigger diabetes (such as amino-acid residues 12 to 20 of the insulin B chain; B:12–20) can bind to a receptor termed I-A^{g7} on an antigen-presenting cell (APC) of the immune system. When such a peptide-bound APC encounters a T cell that recognizes the peptide through its T-cell receptor (TCR), this activates the T cell and sets in motion a process that can cause diabetes.

mutation in insulin that renders the antigen non-immunogenic. Several months later, the authors sequenced RNA from T cells derived from the transferred cells. These T cells expressed genes characteristic of an activated state only in the mice that had the immunogenic version of the antigen. Moreover, when these activated T cells were transferred into NOD mice that lacked T cells, they caused type 1 diabetes much more rapidly than when T cells from the control mice were used instead, confirming the role of these cells and this specific antigen in the generation of diabetes.

The authors sought to investigate how and where the diabetes-causing CD4 T cells are activated. They devised a highly sensitive and technically challenging microscopy approach that measures T-cell recognition of I-A^{g7}–B:12–20 complexes by monitoring a decrease in the mobility of CD4 T cells in mouse lymph nodes transferred from animals into a culture medium that promoted T-cell proliferation. This approach builds on previous findings that presentation of this antigen occurs in lymph nodes throughout the body, not just in those that drain from the pancreas[7]. Control experiments established that CD4 T cells are not activated by this antigen in NOD mice lacking I-A^{g7} or in animals that have an insulin mutation that prevents B:12–20 peptide binding to I-A^{g7}.

How do antigen-presenting cells in the lymph node acquire insulin peptides? One possibility is that they take up and process full-length insulin, given that these cells express insulin-binding receptors. However, when mice received a drug that blocks insulin binding and uptake, this did not block insulin recognition by CD4 T cells in lymph nodes.

So where are these insulin peptides generated? The authors turned their attention to pancreatic β-cells. These cells contain large amounts of insulin stored in granules, which are released into the bloodstream when blood glucose rises after a meal. As a quality-control measure, insulin has an 'expiration date', and old insulin granules are 'retired' and degraded in an organelle called a crinosome, which forms when a granule fuses with an organelle called a lysosome (Fig. 1).

By using a mass-spectrometry technique to identify peptides, Wan *et al.* found that crinosomes, but not insulin granules, contain substantial amounts of insulin peptides that are associated with diabetes. The authors determined that intravenously administered insulin peptides can rapidly reach antigen-presenting cells in lymph nodes, consistent with a model in which insulin peptides linked to diabetes are released into the bloodstream.

By providing an explanation of how disease-causing T cells can be activated, Wan and colleagues' findings raise many important questions. Given the short life expectancy for patients with untreated type 1 diabetes, one might have expected robust selection pressure against HLA-DQ8. Yet it is present at extremely high frequency in many ethnic groups[8], suggesting that there is a strong counterbalancing selective advantage in retaining HLA-DQ8. And why isn't there selection against the generation of secreted insulin peptides that trigger autoimmunity in humans? Perhaps the peptides have some positive functions, such as hormonal activity, that might also explain the timing of their release from β-cells together with insulin on glucose stimulation. It will be interesting to learn whether different types of secretory cell release peptides generated in crinosomes, and whether this contributes to other autoimmune diseases.

Wan *et al.* show that the activation of diabetes-causing T cells in NOD mice is independent of antigen presentation by B cells or dendritic cells. Which type of antigen-presenting cell is therefore responsible? Perhaps the most puzzling question of all is how insulin peptides activate T cells in the absence of the inflammatory signals that are usually needed to trigger an immune response. NOD mice are maintained in pathogen-free conditions, so they are unlikely to have infections that could provide the necessary inflammatory cues. Microorganisms that naturally reside in NOD mice could potentially provide immune triggers. Indeed, antibiotics can modulate type 1 diabetes susceptibility in NOD mice[9], implicating bacteria in disease pathogenesis.

Although insulin supplements can provide a profound improvement in diabetes treatment, the situation is far from perfect and the life expectancy of those who develop type 1 diabetes is decreased by more than a decade[10]. By uncovering key steps in the development of this condition, Wan and colleagues' findings might hasten the day when type 1 diabetes is relegated to medical history. ■

Jiajie Wei *and* **Jonathan W. Yewdell** *are in the Laboratory of Viral Diseases, National Institutes of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA.*
e-mail: jyewdell@mail.nih.gov

1. Wan, X. *et al. Nature* **560**, 107–111 (2018).
2. Banting, F. G. *et al. Can. Med. Assoc. J.* **12**, 141–146 (1922).
3. Wan, X. & Unanue, E. R. *Curr. Opin. Immunol.* **46**, 30–37 (2017).
4. Wallberg, M. & Cooke, A. *Trends Immunol.* **34**, 583–591 (2013).
5. DiMeglio, L. A., Evans-Molina, C. & Oram, R. A. *Lancet* **391**, 2449–2462 (2018).
6. Suri, A., Walters, J. J., Gross, M. L. & Unanue, E. R. *J Clin. Invest.* **115**, 2268–2276 (2005).
7. Wan, X., Thomas, J. W. & Unanue, E. R. *J. Exp. Med.* **213**, 967–978 (2016).
8. Middleton, D., Menchaca, L., Rood, H. & Komerofsky, R. *Tissue Antigens* **61**, 403–407 (2003).
9. Pearson, J. A., Wong, F. S. & Wen, L. *J. Autoimmun.* **66**, 76–88 (2016).
10. Livingstone, S. J. *et al. J. Am. Med. Assoc.* **313**, 37–44 (2015).

This article was published online on 18 July 2018.

© 2018 Springer Nature Limited. All rights reserved.

EARTH SCIENCE

Origin of the blues



PETRA DIAMONDS LIMITED

Some diamonds are a beautiful blue colour, because they contain trace quantities of boron. But blue diamonds are formed in Earth's mantle, whereas boron is concentrated in the crust. So where did these diamonds get their boron from? On page 84, Smith *et al*. provide the answer to this geochemical conundrum (E. M. Smith *et al*. *Nature* **560**, 84–87; 2018).

The authors analysed minerals trapped in 46 blue diamonds, and from this worked out that the gemstones must have formed in the lower mantle. Their analysis also suggests that the material from which the diamonds formed contained water and came from the oceanic lithosphere (tectonic plates beneath the sea), a rich source of boron.

The findings mean that blue diamonds are some of the deepest ever found. Moreover, they reveal a geochemical pathway that extends from the oceanic lithosphere at Earth's surface to the lower mantle, and a potential route for the ultra-deep cycling of water in our planet. **Andrew Mitchinson**

MEDICAL RESEARCH

A cancer shortcut to the nervous system

Certain cancers are prone to invade the nervous system, which leads to poorer prognosis. A study of leukaemia in mice reveals an unexpectedly direct invasion route from the bone marrow to the central nervous system. SEE ARTICLE P.55

FRANK WINKLER

If malignant cells from solid or blood cancers enter the central nervous system (CNS) and grow there, the treatment options and clinical outlook deteriorate rapidly. In a type of leukaemia called acute lymphoblastic leukaemia (ALL), invasion of the CNS commonly occurs. To try to limit this, people with the condition often receive radiation or chemotherapy that targets the CNS. If more-effective and less-toxic approaches became available to prevent disease spread to the CNS, this might benefit many people with ALL. On page 55, Yao *et al*.[1] report a hitherto unknown route that ALL cells use to enter the CNS, and suggest a possible therapeutic approach that is worth investigating.

When leukaemia spreads into the CNS, this process, termed metastasis, is mainly limited to the region known as the subarachnoid space, which contains cerebrospinal fluid that bathes the brain and spinal cord. The subarachnoid space is surrounded by membranes called the dura mater, the pia mater and the arachnoid, which are collectively called the meninges and are also colonized by cancer cells (Fig. 1). Yao and colleagues used mouse models of ALL to investigate how human leukaemia cells spread into the CNS. They focused on the enzyme PI3K, which is a key regulator of signalling pathways needed for growth, survival and invasion by cancer cells.

The authors studied three different mouse models of ALL, and gave animals a molecule called GS-649443 that inhibits a version of PI3K called PI3Kδ. Animals that received the drug had reduced signs of cancer invasion of the CNS compared with those that did not receive it. The authors studied bone-marrow sites that are commonly rich in leukaemic cells, and found no evidence that the drug was affecting the growth or motility of leukaemic cells or cancer progression at sites outside the CNS. This result suggests that PI3K inhibition specifically affects the ability of leukaemic cells to enter the CNS — which is surprising, because the blood–brain barrier usually poses a formidable obstacle that restricts the ability of molecules or cells to exit blood vessels and enter the brain or the cerebrospinal fluid. Cancer cells that reach the CNS are likely to be beyond the reach of drugs present in the body's bloodstream.

Yao and colleagues analysed gene expression in cancer cells to identify genes whose expression decreased as a result of PI3K inhibition. This pointed them towards the gene that encodes a receptor protein called α6 integrin. This receptor can bind a protein called laminin, which is a key component of the extracellular material that surrounds large blood vessels. Laminin is also a component of the meninges[2] and is present in a CNS structure called the choroid plexus, which is where

© 2018 Springer Nature Limited. All rights reserved.

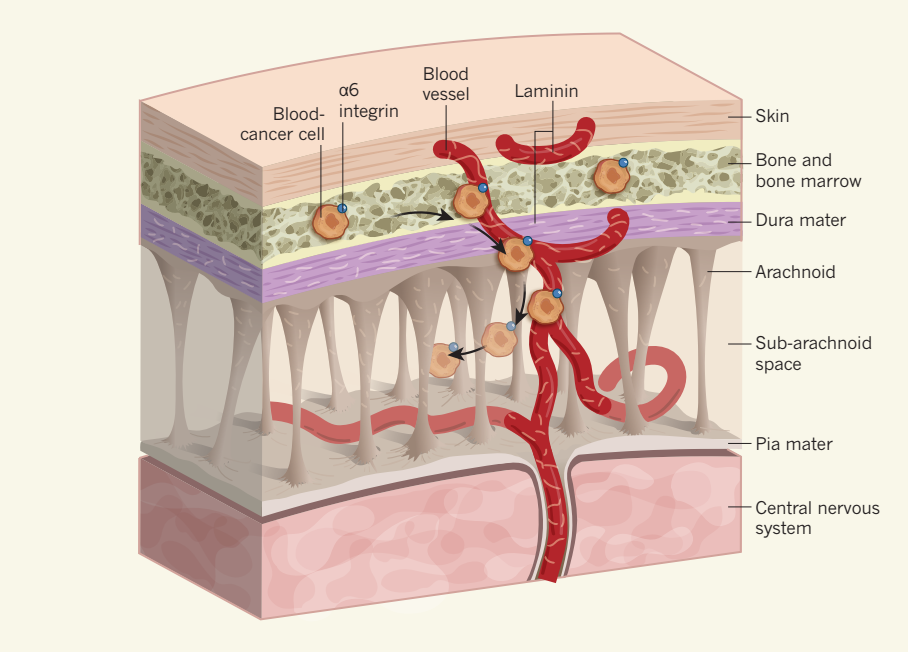


Figure 1 | A route for cancer entry into the central nervous system. Cancer migration to the central nervous system (CNS, which comprises the brain and spinal cord) is often associated with poor prognosis. Yao *et al.*[1] report that human blood-cancer cells in mice reach the central nervous system by migrating along the external surface of blood vessels. This solves the mystery of how these cells leave their site of origin in the bone marrow and reach a region called the subarachnoid space, which contains cerebrospinal fluid. This is located in a region termed the meninges, which covers the CNS and contains membrane layers called the dura mater, arachnoid and pia mater. The authors report that this migration process depends on an enzyme called PI3Kδ (not shown), and requires a receptor protein called α6 integrin on cancer cells. This receptor can bind to the protein laminin, which coats the surface of blood vessels and is also found in the meninges. Laminin aids the migration of neuronal cells during development[2], so perhaps these cancer cells have hijacked components of a natural migration process.

immune cells and cells from solid tumours often enter the CNS[3].

When solid brain tumours arise from a cancer that has spread from elsewhere in the body, a common theme is tumour-cell entry into the brain by a process that requires signalling by a type of integrin protein called β1 integrin, which binds to laminin-containing blood vessels deep in the brain[4,5]. However, when Yao and colleagues analysed brain cells using microscopy techniques, they found no evidence that the choroid plexus or these vessels in the brain are places where ALL cells enter the CNS, confirming the results of previous human and mouse studies[6].

Laminins and laminin-binding integrin receptors act in neuronal path-finding processes during development[2], and are also key mediators of migration processes for healthy and cancerous cells[5,7]. Using microscopy approaches in their mouse models, Yao and colleagues discovered that small blood vessels coated with laminin that connect the bone marrow of the skull and the nearby meninges are sites with high levels of ALL cells. Transit of ALL cells along the external surface of these vessels could provide a direct route for the cells to reach the CNS. Previous studies[6,8] have led to speculation about whether ALL cells spread to the CNS by a direct route, given that the outer layer of the meninges — the dura mater — can have a high level of infiltration by ALL cells.

Conducting studies *in vitro*, the authors found that cerebrospinal fluid contains chemokine molecules that can provide a chemical attractant for ALL cells. Yao and colleagues' work also indicates that ALL-cell migration towards laminin depends on the presence of α6 integrin and can be blocked by PI3Kδ inhibition. The authors observed that treating mice with GS-649443 consistently diminished the invasion of ALL cells along the vascular corridors between the bone marrow and CNS. Moreover, giving animals an antibody that blocks α6 integrin prevented ALL spread to the CNS without having strong effects on ALL progression outside the CNS. Finally, the authors carried out a limited investigation of clinical samples from 26 people with ALL and found that a higher level of expression of α6 integrin correlated with a higher probability that a person's cancer had invaded the CNS.

Interestingly, the ability of cancer cells to hijack pathways required for neurodevelopment has been demonstrated for key steps leading to brain-tumour progression[9], which suggests that this might be a key mechanism underlying how cancer cells successfully colonize the CNS.

Efforts to prevent cancers from reaching the CNS might be a particularly effective approach for improving the clinical outlook for many of these diseases. In a mouse model[4] of tumour progression to brain sites, inhibition of the protein VEGF-A hinders a crucial early switch needed for the formation of tumour blood vessels. In ALL, VEGF-A expression levels are high, and a VEGF-A blocking antibody can prevent the disease invading the meninges of the CNS[10]. Yao and colleagues found that PI3Kδ inhibition results in substantially lower expression of VEGF-A (see Extended Data Fig. 4b of the paper[1]), raising the question of whether this protein is also involved in the colonization of the CNS in ALL. Because blood-vessel formation is not likely to be involved for ALL, given that it is not a solid cancer, perhaps VEGF-A alters the blood vessels that penetrate the CNS, to make them more likely to provide routes for ALL invasion.

Questions remain about the exact route of ALL entry to the CNS. Do the microvessels that directly bridge this route and penetrate tiny holes in the skull, as demonstrated by Yao and colleagues in their mouse studies, also exist in humans? If cancer cells follow the surfaces of known larger blood vessels to travel from the bone marrow to the subarachnoid space in humans, such a route would be of considerable anatomical complexity. Moreover, cancer cells from solid tumours regularly colonize the bone marrow and frequently invade bone, so it would be interesting to determine whether such cells can also use the route taken by ALL to reach the subarachnoid space.

The presence of laminin at all known sites of cancer-cell entry to the CNS, and the ability of laminin exposure to boost the survival of cancer cells[3,4] clearly indicates that more-extensive studies to investigate laminin's role in cancer-cell entry to the CNS are called for. Which versions of integrins and laminin are involved, and which exact routes are used by cancer cells? Finally, Yao and colleagues' insights could be relevant to the clinic. Another PI3Kδ inhibitor, called idelalisib, is already in use to treat certain blood cancers. Could this drug finally offer a way to target the CNS entry of ALL or other types of cancer? ■

Frank Winkler *is in the Department of Neurology, University Hospital Heidelberg, and is also at the German Cancer Research Center, 69120 Heidelberg, Germany.*
e-mail: frank.winkler@med.uni-heidelberg.de

1. Yao, H. *et al. Nature* **560**, 55–60 (2018).
2. Colognato, H., ffrench-Constant, C. & Feltri, M. L. *Trends Neurosci.* **28**, 480–486 (2005).
3. Boire, A. *et al. Cell* **168**, 1101–1113 (2017).
4. Kienast, Y. *et al. Nature Med.* **16**, 116–122 (2010).
5. Carbonell, W. S., Ansorge, O., Sibson, N. & Muschel, R. *PLoS ONE* **4**, e5857 (2009).
6. Thomas, L. B. *Cancer Res.* **25**, 1555–1571 (1965).
7. Gritsenko, P. G., Ilina, O. & Friedl, P. *J. Pathology* **226**, 185–199 (2012).
8. Williams, M. T. S. *et al. Blood* **127**, 1998–2006 (2016).
9. Osswald, M. *et al. Nature* **528**, 93–98 (2015).
10. Münch, V. *et al. Blood* **130**, 643–654 (2017).

This article was published online on 18 July 2018.

© 2018 Springer Nature Limited. All rights reserved.

ACOUSTICS

Negative refraction without reflection

At the interface between two facets of an artificial crystal, sound waves can be transmitted in the opposite direction to that expected, and undergo no reflection. Such wave behaviour could have many applications. SEE LETTER P.61

BAILE ZHANG

Waves change direction when they pass from one medium to another — a phenomenon called refraction. This effect underlies most optical lenses and instruments, and is widely found in acoustics when an acoustic beam behaves like an optical beam. In general, some of the waves are reflected during the refraction process. On page 61, He *et al.*[1] report an impressive demonstration of a previously unobserved refraction phenomenon. They show that, in a certain artificially engineered material, an acoustic beam can be refracted in the opposite direction to that seen in ordinary materials, without reflection. The authors' findings could lead to improved control of waves in electronic and photonic systems.

When an acoustic or optical ray strikes the interface between two different media, part of its energy passes through the interface to form a refracted ray (Fig. 1a). The remaining energy reflects from the interface to produce a reflected ray. In nature, the incident and refracted rays are always on opposite sides of the normal — an imaginary line perpendicular to the interface. But, in theory, this need not be the case.

In 1968, the Russian physicist Victor Veselago considered a hypothetical material that has a negative refractive index[2]. A refractive index describes how waves propagate in a medium, and is positive in all conventional materials. Veselago showed that the way in which refraction usually occurs could be reversed in a negative-index material: the refracted ray could emerge on the same side of the normal as the incident ray (Fig. 1b).

Although intriguing, negative refraction did not trigger much attention, and was considered impossible for more than 30 years because it was thought that negative-index materials could not exist. The situation changed in 2000, when the British physicist John Pendry made a shocking prediction[3]: that negative refraction could be used to make a lens that could focus light more tightly than is normally possible. He also identified a practical way to construct negative-index materials in the lab using artificial structures. Such materials, now generally referred to as metamaterials, stimulated research into concepts such as invisibility cloaking[4] that had previously existed only in science fiction.

In the years since Pendry's work, the pursuit of negative refraction has led to developments in optics, acoustics, plasmonics (the study of how light interacts with electrons in metals) and even graphene-based electronics[5]. Versions of negative refraction have been realized in each of these areas. However, the phenomenon is generally accompanied by reflection, which is often undesirable. In many cases, such as in experiments involving the refraction of electrons through an interface[5], reflection can even dominate negative refraction.

The property of reflection immunity is not found in natural optical materials for light. However, it does occur in exotic phases of matter known as topological quantum matter, for quantum-mechanical electronic waves. A well-studied example is the topological insulator, which is an electrical insulator in its interior, but conducts electricity on its surface through electronic waves called topological surface states. Such states are able to propagate unidirectionally — they bypass obstacles and defects, rather than being reflected.

He and colleagues' demonstration was directly inspired by another emerging topological quantum matter: the Weyl semimetal[6]. The topological surface states in this material cannot propagate in all directions; propagation is confined to a certain range of directions, which connect to form what are known as Fermi arcs[6]. Because the limited range of propagation directions does not include the direction in which reflection would normally occur, reflection is forbidden (Fig. 1c).

In their experiment, He *et al.* used an artificial crystal that is an acoustic analogue of the Weyl semimetal. They found that, at the interface between two adjacent facets of the crystal, airborne acoustic waves could undergo negative refraction without reflection. The authors' results represent the first realization of negative refraction for topological surface states.

There are a few limitations of the work. For instance, the refraction does not occur in a flat plane, contrary to the common impression of refraction. Moreover, the interface scatters some of the acoustic waves into the crystal's interior, resulting in energy loss. Nevertheless, the demonstration opens the door to many exciting opportunities for further research.

The immediate question is whether He and colleagues' refraction phenomenon could be realized in optical systems for light and condensed-matter systems for electrons. Another question, which will be of interest to both optical and condensed-matter physicists, is how to engineer the range of propagation directions — and, in turn, the Fermi arcs — to achieve greater control of negative refraction. In this sense, the authors' work provides the first practical use of Fermi arcs, which are currently being enthusiastically explored in condensed-matter systems[7,8] and in optical structures called photonic crystals[9].

The refraction phenomenon could also find widespread use in acoustics. For example, the combination of negative refraction and zero reflection could lead to improved resolution in ultrasonic imaging and testing. Moreover, acoustic waves are used in biomedical microfluidic devices to trap, sort and deliver cells and drug particles. Reflection-free acoustic waves are strongly desirable in such applications, because reflections at the interfaces and sharp corners of microfluidic channels are currently a huge limitation to device efficiency.

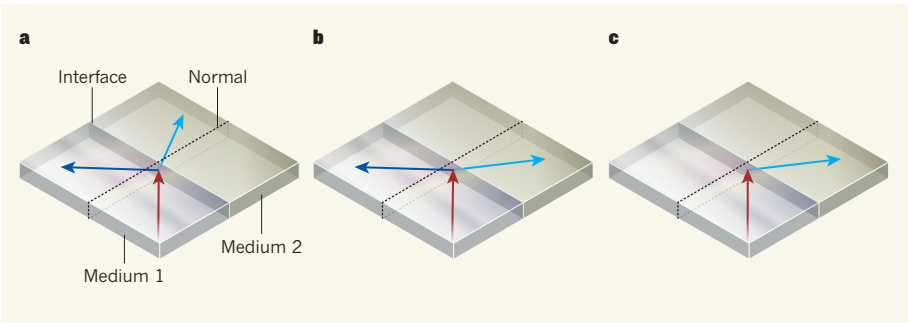


Figure 1 | Comparison of refraction phenomena. **a**, In conventional refraction, when an acoustic or optical ray (red) hits the interface between two different media, a reflected ray (dark blue) and a refracted ray (light blue) are produced. The incident and refracted rays exist on opposite sides of the normal — an imaginary line perpendicular to the interface. **b**, In negative refraction, the refracted ray emerges on the same side of the normal as the incident ray. **c**, He *et al.*[1] report a previously unobserved type of refraction for acoustic rays, in which not only are the incident and refracted rays on the same side of the normal, but also there is no reflected ray. (Figure adapted from ref. 1.)

© 2018 Springer Nature Limited. All rights reserved.



50 Years Ago

The prospect of beer by the litre for British drinkers came a step closer last week. Mr Anthony Wedgwood Benn, the Minister of Technology, announced that the United Kingdom is to adopt the metric system of measurement by 1975 — the target date already accepted by British industry for its timetable … The industrial change is going ahead fast … but the non-industrial sector of the economy and the general public have been lagging behind … It is "imperative" for the planning of the change in the general sectors of the economy to be put in hand; if this is not done, "the dynamism of the industrial change will be lost" … The cost-effectiveness of metrication is not … likely to be known until it is a fact. So far, certainly, hunch has played a greater part than has sober analysis of the situation.

From *Nature* 3 August 1968

100 Years Ago

… Finally, there is the personal use of insecticidal preparations as aids to the primitive method of getting rid of [lice] — now referred to as "chat"-hunting … [T]he preparation should be of quick action and easy of application to clothing, and its issue should be as general and comprehensive as that of food … [P]astes are more economical and convenient than powders; fluids are out of the question. Crude "unwhizzed" naphthalene, produced by coke-oven plants, affords the most effective base, and may be conveniently mixed into paste form by the addition of soft soap or some grease, such as vaseline, in the proportion of 10 to 20 per cent … When it is necessary to use an anti-lice preparation on a hair-clad surface the use of vaseline, to which has been added ½ per cent. of veratrine dissolved in 5 per cent. of benzene, may be recommended.

From *Nature* 1 August 1918

Topological acoustics is therefore a promising research field that not only can produce phenomena that are difficult to realize in other physical systems, but could also bring about transformative technologies. ■

Baile Zhang *is in the Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore.*
e-mail: blzhang@ntu.edu.sg

1. He, H. *et al. Nature* **560**, 61–64 (2018).
2. Veselago, V. G. *Sov. Phys. Usp.* **10**, 509–514 (1968).
3. Pendry, J. B. *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
4. Pendry, J. B., Schurig, D. & Smith, D. R. *Science* **312**, 1780–1782 (2006).
5. Chen, S. *et al. Science* **353**, 1522–1525 (2016).
6. Wan, X., Turner, A. M., Vishwanath, A. & Savrasov, S. Y. *Phys. Rev. B* **83**, 205101 (2011).
7. Xu, S.-Y. *et al. Science* **349**, 613–617 (2015).
8. Lv, B. Q. *et al. Phys. Rev. X* **5**, 031013 (2015).
9. Lu, L. *et al. Science* **349**, 622–624 (2015).

METABOLISM

An unexpected trigger for calorie burning

The molecule succinate, which is a product of metabolism, promotes heat production and therefore calorie burning in brown fat in mice. This discovery could have implications for combating obesity in humans. SEE LETTER P.102

SHENG HUI & JOSHUA D. RABINOWITZ

There are two ways to lose weight: eat less to reduce the number of calories available for metabolism by the body, or burn more calories, for example through exercise. On page 102, Mills *et al.*[1] identify a molecule produced during nutrient metabolism that, surprisingly, induces calorie burning. This metabolite, succinate, activates energy expenditure in brown fat. Remarkably, supplementing the drinking water of mice with succinate can prevent the animals from gaining weight.

Brown fat is different from the white fat that builds up around our waistlines. Whereas white fat acts as an energy reserve, brown fat specializes in heat generation, and is essential for mammals to maintain their body temperature in the cold[2]. Brown-fat cells contain smaller lipid droplets than do white-fat cells, and have many more organelles called mitochondria[3], which enable brown fat to generate heat.

In mitochondria, a metabolic pathway called the TCA cycle breaks down nutrients such as glucose, lactate and fat into carbon dioxide, using the energy stored in the nutrients to generate high-energy electrons. These electrons are used to pump protons (hydrogen ions, H^+) out of the interior matrix of the mitochondrion into the space between the organelle's inner and outer membranes, thereby converting the energy into a proton gradient. Normally, protons re-enter the mitochondrial matrix through a membrane-spanning protein complex called the proton ATPase. This complex uses the energy stored in the proton gradient to convert ADP molecules into energy-carrying ATP molecules, and thereby generates most of the body's usable energy. But in brown fat, protons pass through another protein, uncoupling protein 1 (UCP1). This transporter uncouples the process of crossing the mitochondrial membrane from that of ATP production, effectively wasting the proton gradient's energy as heat (reviewed in ref. 4).

This capacity of brown fat to dissipate calories as heat has attracted much attention, in the hope of activating the process to combat obesity[5]. To do this, it is necessary to know what switches on calorie burning by brown fat. At the macroscopic level, the main answer is exposure to cold. At the mechanistic level, it has been proposed that the brain senses cold and sends signals to brown fat through a process mediated by proteins called β-adrenergic receptors[2]. But drugs that activate these receptors have not been successful in curbing obesity[6]. Thus, there is intense interest in finding new pathways that activate heat generation in brown fat.

Mills *et al.* began by searching for metabolites that are selectively abundant in brown fat, and whose concentration in this tissue increases during cold exposure. Their survey identified succinate, one of the metabolic intermediates of the TCA cycle.

The TCA cycle is generally assumed to be a cell-intrinsic process in which most intermediates are trapped in the mitochondrial matrix. Thus, most succinate is consumed by the same cell that produces it. Some succinate, however, makes its way into the bloodstream. The authors provide evidence that a key trigger for the release of succinate may be muscle activity, because shivering in response to cold increased blood succinate levels in mice.

To trace the fate of succinate circulating in

© 2018 Springer Nature Limited. All rights reserved.

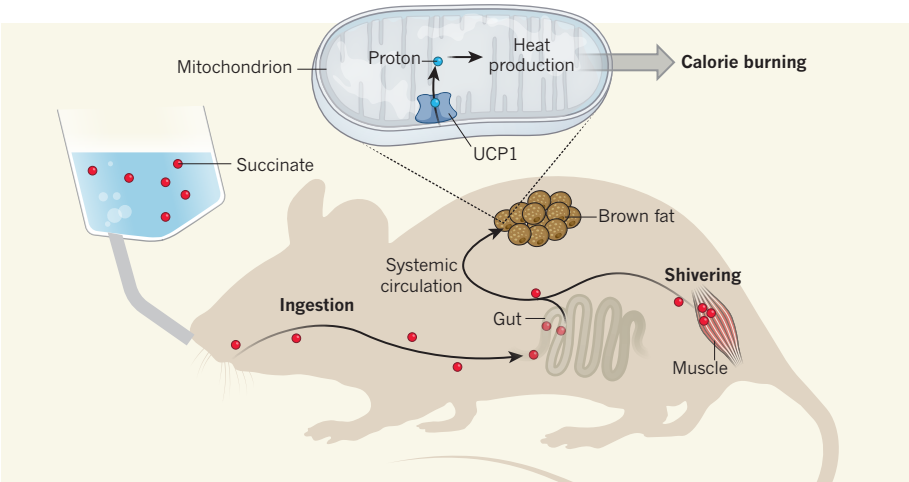


Figure 1 | Circulating succinate molecules mediate calorie burning. Mills *et al.*[1] report a mechanism by which weight gain can be controlled in mice. Succinate added to drinking water is ingested and enters the systemic blood circulation from the gut. Succinate is also released into the circulation by muscle cells during shivering in response to cold. From the circulation, succinate can enter brown-fat cells, which contain many organelles called mitochondria. Here, it triggers the mitochondrial protein UCP1 to leak protons (hydrogen ions; H^+), converting chemical energy into heat and thereby burning calories.

the blood, Mills and colleagues injected mice with succinate tagged by a heavy isotope of carbon. They found that the carbon isotope accumulated preferentially in brown fat. Thus, brown fat seems to be programmed to use circulating succinate as a fuel. Consistent with this, the authors showed that isolated brown-fat cells, but not most other cell types tested, avidly took up and burnt succinate.

Mills *et al.* next showed that acute succinate administration in mice raised the local temperature of brown fat. And, strikingly, administering succinate in drinking water for four weeks prevented obesity in mice on a high-fat diet. These metabolic effects depended on UCP1 — most of the beneficial metabolic effects of succinate were absent in mice genetically engineered to lack this protein. Thus, succinate activates heat production and calorie burning in brown fat (Fig. 1).

How exactly does succinate trigger heat production? In the TCA cycle, succinate is consumed by the enzyme succinate dehydrogenase. The activity of this enzyme produces molecules called reactive oxygen species (ROS), which have been proposed to promote heat generation by brown fat[7]. The authors therefore suggest that succinate accumulation induces calorie burning by increasing the activity of succinate dehydrogenase and so increasing ROS levels. However, it is unclear whether the contribution of circulating succinate to the TCA cycle in brown-fat cells is really sufficient to alter ROS levels and heat generation.

As an alternative explanation, perhaps succinate triggers a yet-to-be-discovered signalling system in brown fat. Or perhaps circulating succinate is sensed in a different part of the body, such as the brain, which then signals to brown fat to activate heat production. Defining the mechanism at work is of more than academic interest — it might prove important in determining the ideal dose and schedule for succinate administration in humans, or for identifying pharmacological alternatives to bulk succinate intake. Finding the molecular players involved will be crucial, the most obvious missing protein being the transporter that carries succinate into brown fat.

Humans, of course, differ from mice in many ways. One of the most obvious is our larger body size, which is associated with a lower ratio of body surface area to mass. As a consequence, we are better at staying warm than are mice, but worse at getting rid of heat. It is probably for these reasons that brown fat makes up a much lower percentage of our body mass[8]. Moreover, we lose brown fat as we age. This could limit the extent to which activation of metabolic processes in brown fat can alter calorie expenditure. Accordingly, methods to induce brown-fat properties in existing white fat might be needed as a complementary approach[5]. It will nevertheless be interesting to see whether succinate can induce substantial calorie burning in humans.

Taking a step back, circulating TCA intermediates have not previously been considered as key factors in metabolism. But several TCA intermediates are present in the circulation at substantial levels, and some of them, such as citrate, flow into and out of the bloodstream to a greater extent than does succinate[9]. The finding that circulating succinate has a well-defined, and perhaps even medically important, metabolic role raises the possibility that circulating TCA intermediates will more generally prove to be vital metabolic players. ■

Sheng Hui *and* **Joshua D. Rabinowitz** *are at the Lewis-Sigler Institute for Integrative Genomics, and in the Department of Chemistry, Princeton University, Princeton, New Jersey 08544, USA.*
e-mails: shui@princeton.edu; joshr@princeton.edu

1. Mills, E. L. *et al. Nature* **560**, 102–106 (2018).
2. Cannon, B. & Nedergaard, J. *Physiol. Rev.* **84**, 277–359 (2004).
3. Rosen, E. D. & Spiegelman, B. M. *Cell* **156**, 20–44 (2014).
4. Nedergaard, J., Ricquier, D. & Kozak, L. P. *EMBO Rep.* **6**, 917–921 (2005).
5. Harms, M. & Seale, P. *Nature Med.* **19**, 1252–1263 (2013).
6. Carey, A. L. *et al. Diabetologia* **56**, 147–155 (2013).
7. Chouchani, E. T., Kazak, L. & Spiegelman, B. M. *J. Bio. Chem.* **292**, 16810–16816 (2017).
8. Enerbäck, S. *Cell Metab.* **11**, 248–252 (2010).
9. Hui, S. *et al. Nature* **551**, 115–118 (2017).

This article was published online on 18 July 2018.

NEURODEVELOPMENT

Nascent neurons need nature and nurture

How genetic and environmental factors contribute to the generation of various subtypes of inhibitory neurons called interneurons in the brain is unclear. A study in mice provides new insight into this process.

CHRISTIAN MAYER & GORD FISHELL

The mature brain contains an enormous variety of locally projecting inhibitory neurons known as interneurons. How the brain's precise complement of interneurons is generated during development is a subject of lively debate. At its heart, this question is one of nature versus nurture. Young interneurons are 'born' in a region called the subpallium and undergo a long migration to reach their final positions in the brain's cortex — but it remains unclear how much of an interneuron's mature fate is bestowed by its genetic identity, which is established when the cell stops proliferating, and how much is acquired through nurture during migration. Writing in *Nature Neuroscience*, Lim *et al.*[1] investigate how migration influences cellular identity.

There is evidence to support roles for both nature and nurture in defining the identities of the different classes, types and subtypes of

© 2018 Springer Nature Limited. All rights reserved.

interneuron in the mature brain. Intrinsic gene-expression programs are thought to begin to define interneuron identities in the embryo, and to unfold over a lengthy period of time[2–4]. The expression of certain genes remains conserved in specific types of interneuron from their birth through to adulthood[4], whereas others affect interneuron maturation more transiently during early embryonic development[5]. Such intrinsic processes are thought to cooperate with the interneuron's environment to establish neuronal circuits and brain connectivity in the adult[6]. For example, after arrival in the cortex, neuronal activity affects several aspects of interneuronal development[7–9].

The authors studied the migration pathways of two types of interneuron — one characterized in the adult by expression of the protein somatostatin, the other by expression of the protein parvalbumin. Both types are born in the same general region of the subpallium. Interneurons from this region reach the embryonic cortex predominantly by two migration routes[10]: one that takes them through the marginal zone above the cortex; and one that transits below the cortex, through the subventricular zone.

Does the route taken by an immature interneuron have an effect on the identity of the mature cell it will become? To find out, Lim *et al.* specifically labelled somatostatin- or parvalbumin-expressing interneuron precursors in the marginal zone with a fluorescent protein, and observed the cells' development. The authors found that both populations tend to develop complex projections at the end of their migration through the marginal zone. These projections, called translaminar axons, cross different layers of the cortex. This finding led the researchers to propose that migration through the marginal zone influences the growth and branching of axons through some general mechanism.

To examine this idea, Lim and colleagues performed time-lapse imaging experiments in brain slices grown in culture. Consistent with their hypothesis, interneurons 'abseiled' down into the cortex after travelling through the marginal zone. During this process, most of these cells anchored their nascent axon in the marginal zone like a trailing rope (Fig. 1). Thus, migration route and axonal development seem to be linked for these cells.

Next, the authors investigated the consequences of deleting genes that are expressed in somatostatin-interneuron precursors that migrate through the marginal zone, but not in those that pass through the subventricular zone. The group found that deletion of one such gene, *Mafb,* in these cells results in about a 20% decrease in the fraction of somatostatin-interneuron precursors migrating through the marginal zone. Moreover, those neurons that failed to migrate through the marginal zone lacked their characteristic translaminar projections. Finally, Lim *et al.* isolated migrating interneurons from both routes and transplanted them back to the beginning of the migration path in a cultured brain slice. Slightly more than 60% of the cells from the marginal zone entered the same migration path again, hinting that intrinsic differences between neurons influence which cells take which path.

Taking their results together, Lim and colleagues conclude that, early in development, genetic factors determine what type of interneuron a cell will become, and direct the cell down the appropriate migratory path. However, there is also evidence from the current work for environmental effects on interneuron development.

First, the effects of migration route on axonal branching seem to be largely independent of genetic priming, because somatostatin- and parvalbumin-interneuron precursors are similarly affected. Second, although *Mafb*-deficient cells that fail to migrate through the marginal zone lack their translaminar projections, they do retain other properties characteristic of cells that follow the marginal-zone route. Third, almost 40% of interneurons transplanted into brain slices from the marginal zone picked a different migratory route the second time round. It is therefore likely that stochastic processes are a major part of the distribution of interneurons between migration paths. A balance between predetermined and environmentally specified aspects of interneuron development seems to be emerging.

One explanation for how such a balance might work involves the expression of genes such as *Mafb* in newly born interneurons acting as a virtual 'look-up table'. In this scenario, the expression of intrinsic signals might bias the response of developing neurons to subsequent environmental cues. As such, the combination of early gene expression coupled with later environmental cues might jointly determine the cells' final identity and connectivity.

Determining the influences of environmental cues on particular interneuron populations requires the ability to selectively target those populations. The identification this year[4,11] of genes expressed early in development that are specific to particular interneuron populations promises to provide a way to probe the contributions of early intrinsic and later environmental cues in particular interneuron subclasses. Certainly, the present paper provides strong evidence for how the two aspects of development are linked. ■

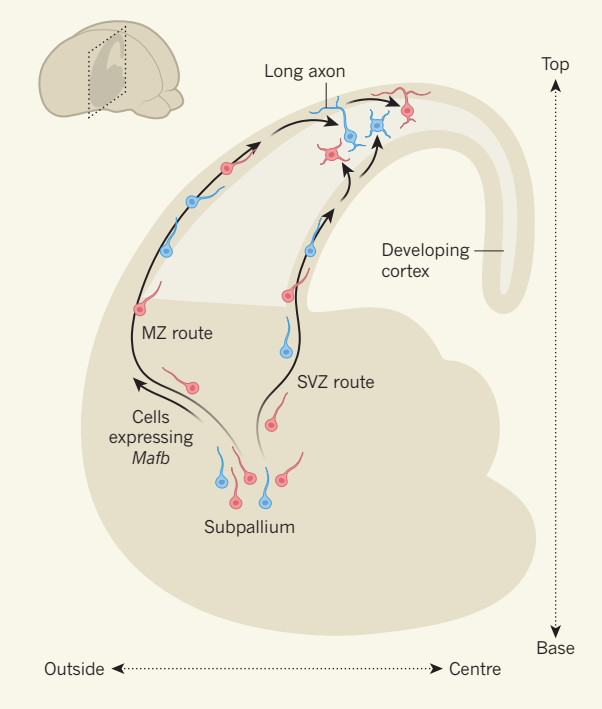


Figure 1 | Intrinsic and environmental cues govern the development of interneurons. Cells called interneurons become progressively more diverse as they mature. Interneurons generated in the subpallium of the mouse brain can be divided into a group that will become mature cells expressing the protein somatostatin (red), and another that will express the protein parvalbumin (blue). Neurons from each group migrate to the cortex through one of two routes: along the marginal zone (MZ) above the developing cortex or along the subventricular zone (SVZ) below it. Lim *et al.*[1] report that an intrinsic cue — expression of the gene *Mafb* — leads cells to migrate through the MZ, and to develop long axonal projections when they move into the cortex, whereas cells that migrate through the SVZ develop short local axons. However, the migration route taken and development of projections also depends on environmental cues, often involving neuronal activity (not shown). The orientation of the brain is indicated in the inset, and by dotted arrows.

Christian Mayer *is at the Max Planck Institute of Neurobiology, 82152 Martinsried, Germany.* **Gord Fishell** *is at the Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA.*
e-mails: cmayer@neuro.mpg.de; gordon_fishell@hms.harvard.edu

1. Lim, L. *et al. Nature Neurosci.* **21**, 920–931 (2018).
2. Wonders, C. P. *et al. Dev. Biol.* **314**, 127–136 (2008).
3. Flames, N. *et al. J. Neurosci.* **27**, 9682–9695 (2007).
4. Mayer, C. *et al. Nature* **555**, 457–462 (2018).
5. Bandler, R. C., Mayer, C. & Fishell, G. *Curr. Opin. Neurobiol.* **42**, 17–24 (2017).
6. Wamsley, B. & Fishell, G. *Nature Rev. Neurosci.* **18**, 299–309 (2017).
7. De Marco García, N. V., Priya, R., Tuncdemir, S. N., Fishell, G. & Karayannis, T. *Nature Neurosci.* **18**, 393–401 (2015).
8. Dehorter, N. *et al. Science* **349**, 1216–1220 (2015).
9. Mardinly, A. R. *et al. Nature* **531**, 371–375 (2016).
10. Tanaka, D. H. & Nakajima, K. *Eur. J. Neurosci.* **35**, 1655–1660 (2012).
11. Mi, D. *et al. Science* **360**, 81–85 (2018).

This article was published online on 23 July 2018.

© 2018 Springer Nature Limited. All rights reserved.

REVIEW

https://doi.org/10.1038/s41586-018-0361-2

Machine learning at the energy and intensity frontiers of particle physics

Alexander Radovic[1]*, Mike Williams[2]*, David Rousseau[3], Michael Kagan[4], Daniele Bonacorsi[5,6], Alexander Himmel[7], Adam Aurisano[8], Kazuhiro Terao[4] & Taritree Wongjirad[9]

Our knowledge of the fundamental particles of nature and their interactions is summarized by the standard model of particle physics. Advancing our understanding in this field has required experiments that operate at ever higher energies and intensities, which produce extremely large and information-rich data samples. The use of machine-learning techniques is revolutionizing how we interpret these data samples, greatly increasing the discovery potential of present and future experiments. Here we summarize the challenges and opportunities that come with the use of machine learning at the frontiers of particle physics.

The standard model of particle physics is supported by an abundance of experimental evidence, yet we know that it cannot be a complete theory of nature because, for example, it cannot incorporate gravity or explain dark matter. Furthermore, many properties of known particles, including neutrinos and the Higgs boson, have not yet been determined experimentally, and the way in which the emergent properties of complex systems of fundamental particles arise from the underlying standard-model theory remains unknown.

Many known particles were discovered using detectors that made subatomic particles visible to the human eye. For example, bubble chambers[1] filled with superheated liquids that boil when charged particles pass through them transform the paths of the particles into visible tracks of bubbles, which can then be photographed and analysed. The detectors at the Large Hadron Collider (LHC)[2] are much more complex and record data at far greater rates than is possible using bubble chambers. For example, the LHCb experiment[3] analyses as many events every six seconds as the Big European Bubble Chamber recorded in its entire 11 years of operation (1973–1983), and the datasets collected by the ATLAS[4] and CMS[5] experiments at the LHC are comparable to the largest industrial data samples. It is impossible for humans to visually inspect such large amounts of data; algorithms running on large computing farms took over this task long ago.

Over the past two decades, particle physics has been migrating towards the use of machine-learning methods in the collection and analysis of its large data samples[6]. Pioneering studies that used neural networks[7,8] and boosted decision trees (BDTs)[9,10] in previous-generation experiments[11–22] laid the groundwork for the emergence of machine learning as an essential tool at the LHC. Machine-learning algorithms made important contributions to the discovery of the Higgs boson[23,24] and most data-analysis tasks now benefit from the use of machine learning. In parallel, the field of machine learning has developed at a rapid pace and, in particular, the subfield of deep learning has delivered superhuman performance in several domains[25,26]. Incorporating these tools while maintaining scientific rigour required in particle-physics analyses presents new challenges. This Review focuses on the application and development of machine-learning methods at the LHC, including recent advances based on deep learning. In addition, we present some example applications of deep learning within the subfield of neutrino physics, in which state-of-the-art methods, such as from computer vision, are naturally applicable.

Big data at the LHC

The sensor arrays of the LHC experiments produce data at a rate of about one petabyte per second. Even after drastic data reduction by the custom-built electronics used to readout the sensor arrays, which involves zero suppression of the sparse data streams and the use of various custom compression algorithms, the data rates are still too large to store the data indefinitely—as much as 50 terabytes per second, resulting in as much data every hour as Facebook collects globally in a year[27]. In this section we first motivate why it is necessary to produce such immense data samples, before discussing how machine learning is being used to more effectively select—in real time—which data to keep for further studies and which data to permanently discard. In addition, we show how the use of machine learning is leading to more efficient processing of these data using vast computing resources distributed around the world. Both of these big-data challenges must be overcome before the LHC data can be used to advance our knowledge of fundamental particles.

LHC operations

Einstein famously related mass m to energy E via $E = mc^2$, where c is the speed of light in a vacuum. A powerful particle accelerator such as the LHC, which is 27 km in circumference, is therefore required to create particles orders of magnitude more massive than the proton, such as the Higgs boson. A Higgs boson is produced only once every few billion proton–proton collisions at the LHC. Many other interesting reactions occur orders of magnitude less often. To enable such data samples to be recorded in a reasonable time frame, the LHC collides nearly one billion protons per second.

High-energy collisions can produce hundreds of particles, and disentangling such complex events requires detectors with large and diverse sensor arrays. The ATLAS and CMS detectors each contain roughly 100 million detection elements. Most of the particles produced in the LHC experiments decay before they can be detected by any of the sensors. Therefore, LHC analyses must infer what the underlying reactions were on the basis of the properties of the particles that are detected. A wide variety of sensor technologies are used in the LHC detectors. The various signals from the particles that are detected by these sensor arrays are digitized, converting the physical processes involving subatomic particles into large collections of bytes.

[1]College of William and Mary, Williamsburg, VA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]LAL, Université Paris-Sud, CNRS/IN2P3, Université Paris-Saclay, Orsay, France. [4]SLAC National Accelerator Laboratory, Menlo Park, CA, USA. [5]Università di Bologna, Bologna, Italy. [6]INFN Sezione di Bologna, Bologna, Italy. [7]Fermi National Accelerator Laboratory, Batavia, IL, USA. [8]University of Cincinnati, Cincinnati, OH, USA. [9]Tufts University, Medford, MA, USA. *e-mail: aradovic@wm.edu; mwill@mit.edu

© 2018 Springer Nature Limited. All rights reserved.

The extreme rate at which the LHC collides protons, along with the size and complexity of the LHC detectors, results in the production of enormous data samples.

Real-time analysis

The LHC experiments use data-reduction schemes executed in real time, referred to as triggers, to identify which data to retain for future analysis and which to permanently discard. For example, the ATLAS and CMS experiments each keep only about 1 in every 100,000 events. Despite this, their data samples are each still about 20 petabytes per year. The first step in deciding which events to keep relies on logic that is encoded directly into the hardware to enable the fastest possible decisions, such as into devices known as field-programmable gate arrays (FPGAs). Machine learning is already used in this environment; for example, CMS uses machine learning in its trigger hardware to better estimate the momentum of muons[28], with the inputs to the algorithm discretized to enable the machine-learning response to be encoded in a large look-up table that is easily programmed into the FPGAs.

In addition, the LHC experiments use huge computing farms to process the extreme volumes of data and search for interesting signatures. In the case of the LHCb experiment, many of the reactions of greatest interest do not produce striking signatures in the detector, making it necessary to thoroughly analyse high-dimensional feature spaces in real time to efficiently classify events[29]. Since the first year of LHCb data collection, the primary algorithm used for such classification has been machine-learning-based; specifically, a BDT was used for the first two years[30], which has since been replaced by a MatrixNet algorithm[31]. The use of machine learning is now ubiquitous, which has greatly improved performance while satisfying the stringent robustness requirements of a system that makes irreversible decisions. Currently, 70% of all data retained are classified by machine-learning algorithms and all charged-particle tracks are vetted by neural networks[32]. As an example of the effect of these machine-learning methods, achieving the same sensitivity as a recent LHCb search for the dark-matter analogue of the photon, which was performed using data collected in 2016[33], would have required 10 years of data collection without the use of machine learning.

Actionable insights from computing metadata

Processing of the industrial-scale data samples collected by the LHC experiments is performed using the computing resources of the LHC Computing Grid, which are distributed across dozens of centres worldwide. The massive volumes of data moved between grid centres, and the large number of CPU processing jobs used to access and analyse these data, generate an enormous amount of metadata information from which actionable insights can be extracted. Machine-learning techniques have recently begun to play a crucial part in increasing the efficiency of computing-resource usage at the LHC[34–36]. One example is predicting which data will be accessed the most, as currently monitored by CMS[37] and LHCb[38], so that it becomes possible to optimize data storage at the grid centres. Another example involves monitoring data-transfer latencies over complex network topologies at CMS[39], using machine learning to identify problematic nodes and to predict likely congestions. Currently, machine learning informs the choices of the computing-operations teams, but in the future it form the basis of fully automatic and adaptive models.

Machine learning as an established tool

After identifying and recording the most interesting LHC events and processing them on the Computing Grid—two vital tasks supported by machine learning—the data are ready for exploration. The first step in interpreting these data involves grouping the signals recorded by various sensor elements according to which particle created them. The types and properties of the particles can then be inferred from the subsets of event information associated with them. Finally, after reconstructing all detected particles in the event, the data are analysed to determine the underlying physical processes that created the particles.

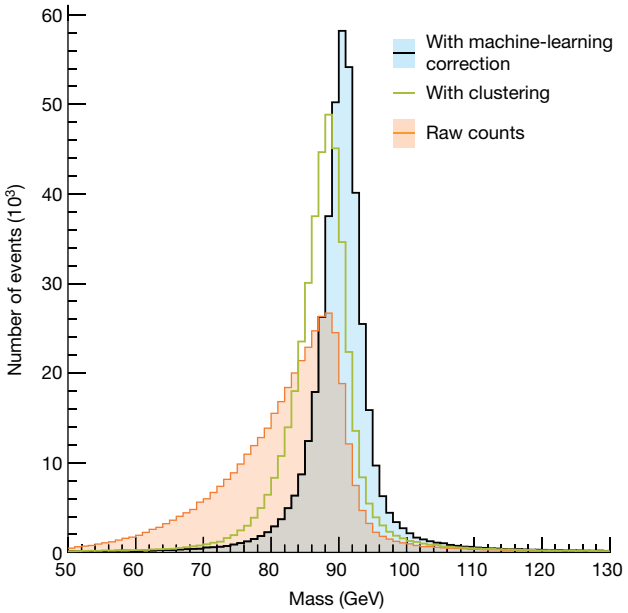


Fig. 1 | Machine learning for calorimetry at CMS. The mass distribution of Z bosons that decay to electron–positron pairs ($Z \rightarrow e^+e^-$), as measured in the central part of the CMS detector and binned into 1-GeV bins, is shown for three cases: using only the raw information from the detector (orange), after clustering the data (green) and after applying the machine-learning-based corrections discussed in the text (blue). The true position of the peak for this decay is 91 GeV. Image adapted from ref. [101] under a CC BY 4.0 license, copyright CERN, reused with permission.

Interpreting such complex data samples is an extremely challenging task, which has been revolutionized by the use of machine-learning techniques. About 2,000 journal articles have been produced by the LHC experiments to date, providing a large library of examples of the use of machine learning with these types of complex dataset. In this section we discuss a few highlights, including the role of machine learning in the discovery of the Higgs boson[23,24].

Determining particle properties

The use of machine learning to improve the determination of particle properties is now commonplace at all of the LHC experiments. For example, BDTs are used to increase the resolution of the CMS electromagnetic calorimeter[40]. When an electron or photon enters such a detector, it rapidly loses its energy, which is subsequently collected and measured by the calorimeter. This deposited energy is often recorded by many different sensors and the readings from these sensors must be clustered together to recover the original energy of the particle. Multivariate regression is used by CMS to train BDTs to provide corrections to these inferred energies on the basis of all of the information contained in each calorimeter sensor. Applying these energy corrections to the decay of a Z boson into an electron–positron pair results in a substantial improvement in mass resolution compared to the traditional clustering approach (see Fig. 1).

Discovery of the Higgs boson

As stated above, a Higgs boson is produced only once every few billion proton–proton collisions at the LHC; however, the Higgs boson usually decays in ways that mimic much more copiously produced processes. The cleanest experimental signature of the Higgs boson involves its decay into two muon–antimuon pairs, which occurs roughly once every 10 trillion proton–proton collisions. This and a few other processes were used in the Higgs discovery analyses. Most were selected owing to their striking experimental signatures, which made it possible to obtain pure signals using relatively simple analyses. An important exception was the analysis of the Higgs boson decaying into two photons by the CMS experiment.

© 2018 Springer Nature Limited. All rights reserved.

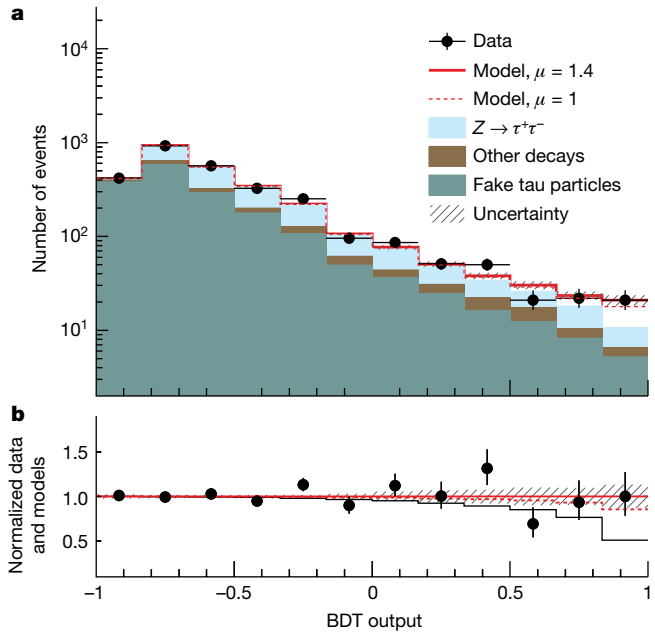


Fig. 2 | Separating signal events from background in the ATLAS experiment. a, The BDT-score distribution for a search for the Higgs boson decaying to a tau-lepton pair ($H \rightarrow \tau^+\tau^-$), with bin widths of 0.17. The black circles show the score of a machine-learning algorithm known as a BDT for data from the ATLAS detector during the 2012 data-taking period, where the error bars show the Poisson statistical uncertainties. This BDT was trained to distinguish a Higgs signal from various non-Higgs backgrounds. The coloured area shows the stacked contributions of the different background processes: $Z \rightarrow \tau^+\tau^-$ decays (blue), other particles decaying to a tau-lepton pair (brown) and fake tau particles (where at least one tau lepton is misidentified; green). The dotted red line shows the expected total counts assuming a Higgs-boson production rate identical to the standard-model expectation ($\mu = 1$); the solid red line shows the expected total counts assuming a Higgs-boson production rate of $\mu = 1.4$, which is still compatible with the standard model; the hatched area shows the systematic uncertainty on the expected total count. The excess counts compared to the coloured region (mainly in the rightmost two bins) are attributed to the Higgs boson. **b**, The ratios of the data (black circles), expected counts for $\mu = 1$ (dashed red line), expected counts for $\mu = 1.4$ (solid red line) and uncertainty (hatched region) from **a** to the expected counts for $\mu = 1.4$ (so the solid red line is identically 1) are shown, along with the ratio of the expected counts excluding the Higgs contribution (the sum of the green, brown, blue and hatched regions in **a**) to the expected counts for $\mu = 1.4$ (black line). Image adapted from ref. [43] under a CC BY 4.0 license, copyright CERN.

The CMS analysis involved searching for a small excess of diphoton candidates, manifested as a narrow peak in the diphoton mass spectrum, in the presence of a large smoothly distributed background. This background largely consisted of diphotons that originated from processes other than the Higgs decay and from candidates formed from one real photon combined with an artificial photon signal (that is, a photon inferred from the detector signals that did not correspond to an actual photon produced in the physical process). Two BDTs were used to improve the diphoton mass resolution by better determining which proton–proton collision the photons were produced in. Because both the standard-model Higgs process and the dominant background processes are well understood, it was possible to use simulated data samples to train a BDT. On the basis of the response of this BDT, the CMS diphotons were either discarded or kept for further analysis. The diphotons selected were also categorized using the BDT response, making it possible to analyse a rare—but highly pure—subset of Higgs decays separately. A simultaneous fit was performed to the mass distributions of all categories, which greatly enhanced the sensitivity to the presence of a Higgs signal. The increase in sensitivity due to the use of machine learning was equivalent to collecting 50% more data.

Determining the properties of the Higgs boson

The standard model contains only one Higgs boson, which is the simplest explanation for the phenomenon known as electroweak symmetry breaking. Many extensions to the standard model predict that there are many Higgs bosons; for example, super-symmetric theories predict a rich Higgs sector, and other theories predict that the Higgs boson is a composite object, not a fundamental particle. The standard model provides precise predictions for the properties of the Higgs boson and it is vital that these predictions are tested experimentally to determine the nature of the Higgs particle discovered at the LHC.

The Higgs-boson discovery analyses firmly established its interactions with the electroweak-force-carrying particles, namely the photon, W boson and Z boson. The standard model also predicts that the Higgs boson interacts with fermions (quarks and leptons) and that the strength of each of these interactions is proportional to the masses of the fermions. This means that the Higgs boson is expected to decay into heavier quarks and leptons more often than into their lighter cousins. The ATLAS and CMS experiments have thus far observed the Higgs boson decaying into the heaviest kinematically accessible quark, the beauty quark[41,42], and into the most massive lepton (a heavier version of the electron known as the tau lepton). Machine learning had a major role in each of these discoveries, although we describe only the ATLAS search for the decay of the Higgs boson into an antitau–tau pair ($H \rightarrow \tau^+\tau^-$) in detail here.

The study of tau particles is challenging because they decay before being detected and because their decays always involve neutrinos that escape detection and carry away energy. Furthermore, the decay $Z \rightarrow \tau^+\tau^-$ occurs about 1,000 times more often than does $H \rightarrow \tau^+\tau^-$. The ATLAS analysis divided the data sample into six distinct kinematic regions. A BDT was trained in each region using 12 weakly discriminating input features[43]. In Fig. 2 we show an example BDT response distribution obtained in one region. The combined analysis of all six regions provided strong evidence for the realization of the Higgs boson coupling to tau leptons in nature, with about 40% better sensitivity achieved through the use of machine learning. Thus far, the interactions of the Higgs boson with quarks and leptons appear to be consistent with standard-model predictions. The simulation that led to this result was eventually released through Kaggle as the basis of the 2014 Higgs Machine Learning Challenge[44], where data scientists competed to provide alternative machine-learning methods to isolate the $H \rightarrow \tau^+\tau^-$ signal. Table 1 shows the impact of machine learning on the measurement of several key processing involving Higgs bosons at ATLAS and CMS.

A high-precision test of the standard model

The standard model predicts that only three out of every billion B_s particles—bound states that contain a beauty quark—decay into a muon–antimuon final state. The fact that this decay rate is so highly suppressed in the standard model contributes to it being extremely sensitive to potential quantum effects induced by as-yet-unknown particles, especially from an extended Higgs sector; for example, certain super-symmetric theories predict an order-of-magnitude enhancement in this decay rate. The CMS and LHCb experiments were the first to find evidence for this decay, using data samples collected in the first few years of the LHC, and a combined analysis of these datasets produced the first observation of it[45]. The analyses used BDTs to reduce the dimensionality of the feature space—excluding the mass—to one dimension and then an analysis was performed of the mass spectra across bins of BDT response. This approach preserved as much information as possible about the mass spectra of both the signal and backgrounds, providing the best possible sensitivity to this extremely rare decay of the B_s meson into a muon–antimuon final state. The decay rate observed is consistent with the standard-model prediction with a precision of about 25%, which places stringent constraints on many proposed extensions to the standard model. Finally, a more recent update from the LHCb experiment achieved the first single-experiment observation[46]; achieving a similar sensitivity without the use of machine

© 2018 Springer Nature Limited. All rights reserved.

Table 1 | Effect of machine learning on the discovery and study of the Higgs boson

Analysis	Years of data collection	Sensitivity without machine learning	Sensitivity with machine learning	Ratio of P values	Additional data required
CMS[24] $H \to \gamma\gamma$	2011–2012	2.2σ, $P = 0.014$	2.7σ, $P = 0.0035$	4.0	51%
ATLAS[43] $H \to \tau^+\tau^-$	2011–2012	2.5σ, $P = 0.0062$	3.4σ, $P = 0.00034$	18	85%
ATLAS[99] $VH \to bb$	2011–2012	1.9σ, $P = 0.029$	2.5σ, $P = 0.0062$	4.7	73%
ATLAS[41] $VH \to bb$	2015–2016	2.8σ, $P = 0.0026$	3.0σ, $P = 0.00135$	1.9	15%
CMS[100] $VH \to bb$	2011–2012	1.4σ, $P = 0.081$	2.1σ, $P = 0.018$	4.5	125%

Five key measurements of three decay modes of the Higgs boson H for which machine learning greatly increased the sensitivity of the LHC experiments, where V denotes a W or Z boson, γ denotes a photon and b a beauty quark. For each analysis, the sensitivity without and with machine learning is given, in terms of both the P values and the equivalent number of Gaussian standard deviations σ. (We present only analyses that provided both machine-learning-based and non-machine-learning-based results; the more recent analyses report only the machine-learning-based results.) The increase in sensitivity achieved by using machine learning, as measured by the ratio of P values, ranges roughly from 2 to 20. An alternative figure of merit is the minimal amount of additional data that would need to be collected to reach the machine-learning-based sensitivity without using machine learning, which varies from 15% to 125%.

learning would have required the collection of about four times as much data. This is just one of many examples of high-precision tests of the standard model at the LHC for which machine learning has markedly increased the power of the measurement.

The emergence of deep learning

Machine learning in particle physics, including the examples presented in the previous two sections, has traditionally involved the use of field-specific knowledge to engineer tools to extract the features of the data that are expected to be the most useful for a given measurement. This enables the incredibly rich initial data to be interpreted using only a small number of features. For example, in the aforementioned B_s decay, a human-designed tracking algorithm first reconstructs the paths taken by the muon and the antimuon in a magnetized particle-physics detector, and from these paths the momenta of the particles are inferred. However, only the dimuon mass and the angle between them are used in the BDT. The rest of the kinematic information is discarded.

For many tasks, information can be lost when these human-designed tools are used to extract features that fail to fully capture the complexity of the problem. As in the fields of computer vision and natural language processing[26,47], there is a growing effort in particle physics to skip the feature-engineering step and instead use the full high-dimensional feature space to train cutting-edge machine-learning algorithms, such as deep neural networks[48]. In this approach, domain expertise is used to design neural-network architectures that are best suited to the specific problem. Studies of such applications have grown substantially in number and complexity within the past several years, beginning around 2014 with applications of deep neural networks to data analysis[49], quickly expanding to the first applications of computer vision[50–52] and to the current broad study of deep learning throughout the field of particle physics[53–56].

In this section we highlight a few recent applications of two types of deep learning algorithm in particle physics: convolutional and recurrent neural networks (CNNs and RNNs, respectively)[57,58]. The outputs of many particle-physics detectors can be viewed as images, and the application of computer-vision techniques is being explored in simplified settings by the LHC community[59–65] and with initial studies on ATLAS and CMS simulations[66,67]. However, such techniques are more naturally applicable in the area of neutrino physics, for which reason we focus our discussion of CNNs to neutrino experiments. Similarly, there are many applications of RNNs, but for brevity we discuss only their use for the study of high-energy beauty quarks at ATLAS and CMS.

Computer vision for neutrino experiments

Loosely inspired by the structure of the visual cortex, CNNs use a strategy that decreases their sensitivity to the absolute position of elements in an image and that makes them more robust to noise. Deep CNNs are able to extract complex features from images and now outperform humans in certain image-classification tasks. Another strength of CNNs is their ability to identify objects in an image, as demonstrated for example by their use in self-driving cars, owing to translation-invariant feature learning. This translational invariance presents a challenge for the LHC experiments, whose detectors consist of layers of distinct detector technologies moving out from the proton–proton collision region. These detectors provide rich information in the absolute reference frame of the detector, which is transformed into a more natural format for a CNN-based approach. By contrast, this characteristic of CNNs is particularly useful for neutrino experiments, which necessarily use large homogeneous detectors owing to the incredibly small probability that a neutrino will interact within a small volume of material. A neutrino interaction can take place anywhere within these detectors and locating them is a critical part of neutrino-physics analyses.

The detectors of the NOvA experiment[68] are filled with scintillating mineral oil, which emits light when a charged particle passes through it. Each NOvA event consists of two images: one taken from the top and the other from the side. The NOvA collaboration has developed a machine-learning algorithm[52] composed of two parallel networks inspired by the GoogleNet[69] architecture. The NOvA CNN extracts features from both views simultaneously and combines them to categorize neutrino interactions in the detector. This network, which improves the efficiency of selecting electron neutrinos by 40% with no loss in purity, has served as the event classifier in searches both for the appearance of electron neutrinos[70] and for a new type of particle called a sterile neutrino[71].

The detector at the MicroBooNE experiment[72], which contains 90 tonnes of liquid argon, detects neutrinos sent from the booster neutrino beamline at Fermilab. Each MircoBooNE event corresponds to a 33-megapixel image that probably contains background tracks caused by cosmic rays. Identifying signals of neutrino interactions in such events, in which both the signal and background tracks vary in size from a few centimetres to metres, is one of the most challenging tasks of the experiment. MicroBooNE recently demonstrated the ability to detect neutrino interactions using a CNN[73]. Specifically, an algorithm called Faster-RCNN[74] uses spatially sensitive information from intermediate convolution layers to predict a bounding box that contains the secondary particles produced in a neutrino interaction. In Fig. 3 we show an example output in which the network successfully localized a neutrino interaction with high confidence. Finally, by taking advantage of accelerated computing on GPUs, these CNNs can run much faster than the conventional algorithms used by previous neutrino experiments. This makes them ideally suited to the task of real-time image classification and object detection.

RNNs for beauty-quark identification

The study of high-energy beauty quarks is of great interest at the LHC because these particles are frequently produced in the decays of Higgs bosons and top quarks and are predicted to be important components of the decays of super-symmetric and other hypothetical particles. A high-energy beauty quark radiates a substantial fraction of its energy in the form of a collimated stream of particles, called a jet, before forming a bound state with an antiquark or two additional quarks. This radiation is emitted over a distance comparable to the size of a proton, making it impossible to observe the emission process directly. The beauty-quark bound states live for only a picosecond, corresponding to millimetre- to centimetre-scale flight distances at the LHC, before randomly decaying into one of a thousand possible sets of commonly produced particles. Therefore, to identify jets that originate from high-energy beauty quarks, it is necessary to be able to determine whether particles were produced directly in the proton–proton collision or in the subsequent decay of a long-lived bound state at a location displaced

© 2018 Springer Nature Limited. All rights reserved.

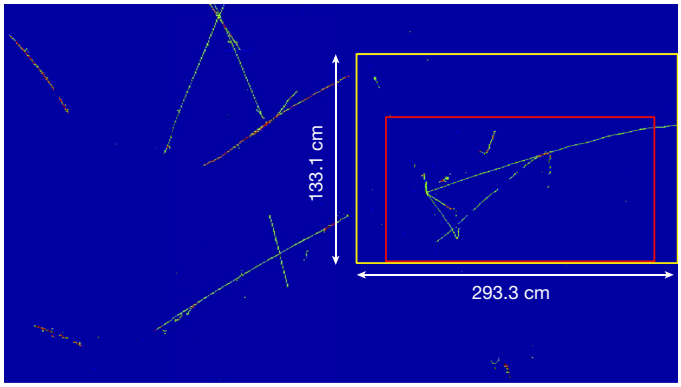


Fig. 3 | Neutrino selection and isolation in MicroBooNE. The MicroBooNE event display shows a simulated neutrino interaction (inside the yellow box) overlaid on a cosmic-ray background image taken using the real detector. The yellow boxed region contains all charge depositions caused by secondary charged particles being produced in the simulated neutrino interaction. The CNN receives as input the display without the yellow box indicated and draws the red box, which matches the yellow box remarkably well and successfully captures the most interesting part of the neutrino interaction. Image adapted from ref. [73], copyright Sissa Medialab, reused with permission of IOP Publishing.

from the proton–proton collision. Because jets typically contain between 10 and 50 particles, the number of potential discriminating features varies on a per-jet basis. Traditional jet-identification algorithms rely on either identifying secondary production points explicitly from the crossing of particle trajectories, a highly challenging task, or compressing the information with engineered features and neglecting the correlations between particles when using single-particle features. Although such algorithms have been combined with machine learning for some time[75,76], machine learning could also be used to improve identification by using the low-level particle features within a jet.

RNNs have proven to be extremely successful at processing long sequences of data, most famously acting as the core of Google's current translation service[47]. RNNs process sequences in such a way that information across the entire sequence can be accumulated and used. Applying an RNN to jet classification requires the particles in the jet to be ordered to form a sequence, such as by ranking them by how incompatible they are with originating from the proton–proton collision. A set of features for each particle is provided to the RNN, which is trained to discriminate between beauty-quark jets and all other types of jet. The use of an RNN at the ATLAS experiment reduced the misidentification rate by a factor of four relative to a non-machine-learning-based algorithm[77]. When the RNN is itself used as an input feature in the subsequent training of a BDT or neural network, the misidentification rate was reduced by a factor of three relative to the machine-learning result without the use of the RNN as an input feature[78]. Similar approaches are also being explored at the CMS experiment[79]; more sophisticated RNN structures have been studied in a simplified setting and show promising results[80].

Training and validation

The machine-learning algorithms used in particle physics are typically trained using supervised learning[81] and data samples for which the true origins, identities and properties of the particles are known a priori. The algorithms learn to identify patterns in the training data, making it possible for them to predict information about particles in data samples for which expert labelling of data is impossible. It is vital that any machine-learning tool undergoes rigorous validation and testing and that the uncertainty on its performance is well understood. There is always the possibility that some features used by a machine-learning algorithm are not properly modelled in the training samples, which—if not properly accounted for—could lead to a false discovery. Ultimately, we use machine-learning tools to minimize uncertainties; the validation procedures discussed in this section are important for gaining confidence in the behaviour of these tools.

Learning from simulation

The need to understand what signals will look like in the detectors and what other processes can mimic the signals has led to the development of high-quality simulation tools. Furthermore, the standard model provides accurate predictions of the rates and kinematic distributions of many of the processes that can mimic interesting signatures (referred to as backgrounds) and that contribute to particle-physics data samples, providing important benchmarks for validating the simulation tools and understanding their uncertainties. Therefore, simulated data samples are often used to train the machine-learning algorithms because in such samples all information is known by construction. An important exception is that it is often possible to obtain highly pure background-only data samples, such as by using events collected under different experimental conditions, and such samples are often used as background samples during training. A hybrid approach is also possible. The MicroBooNE CNN discussed above was trained using simulated neutrino interactions overlaid on cosmic-ray background images taken with the real detector.

Testing for bias

The quality and robustness of all machine-learning tools are validated using well-known reactions recorded by the experiments. One approach, which is used by all LHC experiments[82,83], involves constructing data samples in which the data are fully understood without the use of machine learning. For example, the LHCb experiment uses $J/\psi \rightarrow \mu^+\mu^-$ decays to validate its muon-identification neural network (μNN)[32]; J/ψ is a copiously produced charm–anticharm bound state, which can be identified with 99.9% purity when basing a selection criterion on the μNN response to either the antimuon (μ^+) or muon (μ^-). The identity of the other particle is therefore known without using the μNN, providing an unbiased data sample with which the performance of the μNN can be studied. Domain-specific knowledge is then used to transfer what is learned on these validation samples, in terms of both the expected performance and its uncertainty, to any analysis that uses that specific machine-learning algorithm. In the case of the μNN, the algorithm is studied in ranges of the values of muon and event-level properties and the response of the detector within these ranges is assumed to be independent of the process that produces the muon.

Another approach involves hybrid events, whereby the data are augmented with simulations to produce a test sample that mimics a signal. One example used by NOvA[84] takes abundant and pure muon-neutrino charged-current data and replaces the detected muon with a simulated electron. These hybrid events allow the performance of NOvA's machine-learning algorithms to be studied on rare electron-neutrino charged-current interactions, which are expected to look identical to muon-neutrino charged-current interactions in the detector apart from the muon-to-electron swap. Similar techniques were used for the $H \rightarrow \tau^+\tau^-$ decay by ATLAS[43] and CMS[85].

The approaches presented above are reminiscent of the procedures used to characterize the performance of complex detectors in past decades. Alternatively, tools developed by the machine-learning community can be used to probe the response of the algorithms. For example, t-distributed stochastic neighbour embedding (t-SNE)[86] is a non-parametric embedding technique that allows the proximity of points in a high-dimensional space to be visualized using only two dimensions. It can be used to study the groupings of different events according to the features extracted by a deep neural network. Events with overlapping extracted features, which the network interprets to be similar, are near each other in the t-SNE mapping; conversely, events with little or no overlap are far from each other in the mapping. These t-SNE projections are used to ensure that the groupings match the intuition about the physical processes being studied, to check whether non-training events are embedded as expected, and can even be used in conjunction with auto-encoder neural networks to search for anomalies

© 2018 Springer Nature Limited. All rights reserved.

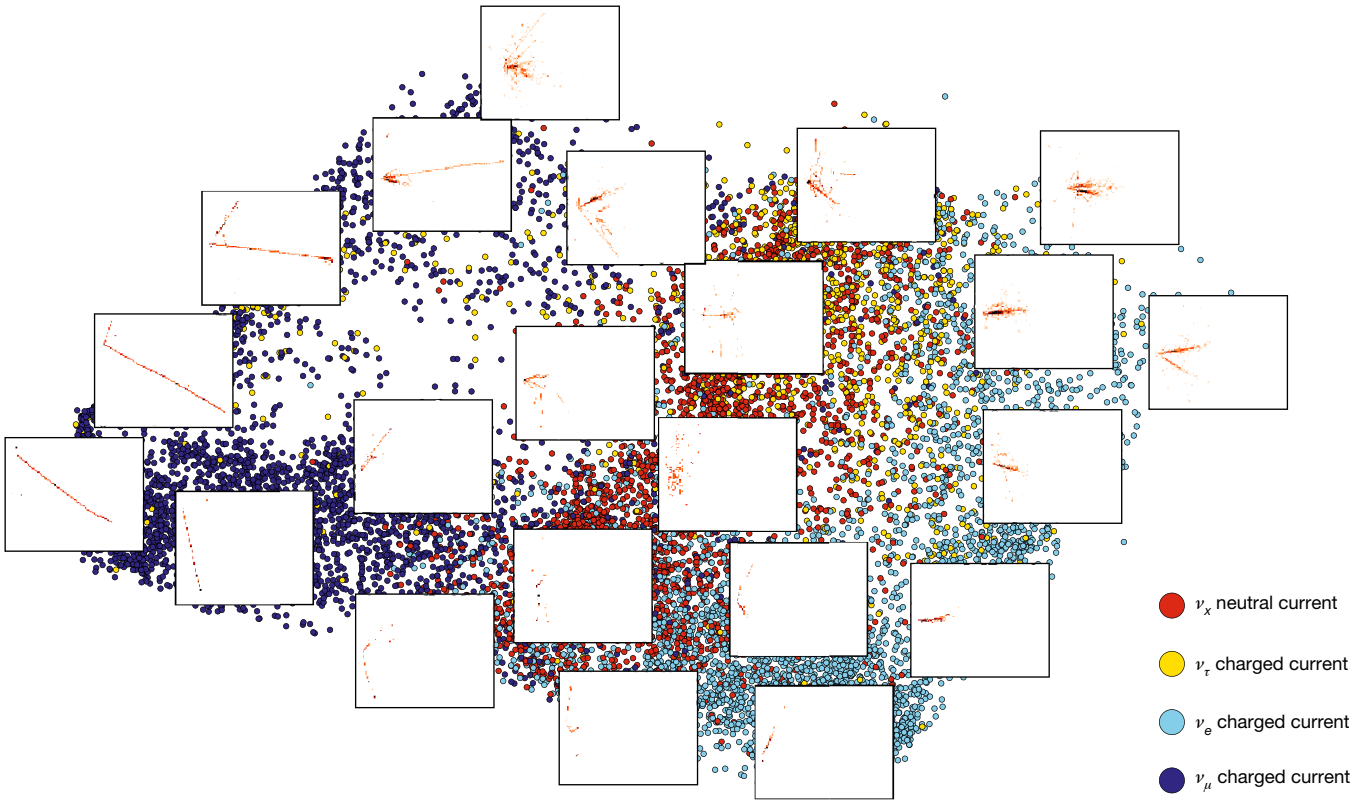


Fig. 4 | Exploring NOvA's event-selection neural network using *t*-SNE. The features extracted using NOvA's neutrino-interaction CNN are projected into a two-dimensional space using the *t*-SNE method. The points represent events from the CNN training sample, with the colours denoting the true event types: muon-neutrino (ν_μ) charged-current interactions (dark blue), electron-neutrino (ν_e) charged-current interactions (light blue), tau-neutrino (ν_τ) charged-current interactions (yellow) and various neutrino (ν_x) neutral-current interactions (red). The subplots show example event topologies from points in the two-dimensional *t*-SNE space, with the intensity of the colour indicating the amount of energy deposited and the axes denoting the spatial location of the charge deposits in the detector. The various types of event are clustered into distinct regions in the horizontal direction, while the multiplicity of the particles in each event is found to be correlated with the location of the events in the vertical direction.

in large datasets. In Fig. 4 we provide an example of such a *t*-SNE using simulated neutrino interactions at the NOvA experiment.

Conclusions and outlook

Within the next decade the LHC will increase the rate at which it collides protons by an order of magnitude, resulting in much higher data rates and even more complex events to disentangle. Neutrino-physics detectors will continue to increase in size and complexity. The tasks discussed in this Review will become even more challenging. Fortunately, machine learning is advancing rapidly, producing tools that are potentially applicable to a wide array of tasks in particle physics. By continuing to map the challenges faced in particle physics to those addressed by the machine-learning community, it is possible to turn the latest developments in machine learning into tools for discovery in high-energy particle physics, such as by conducting machine-learning competitions with LHC benchmark datasets (https://www.kaggle.com/c/trackml-particle-identification). We briefly discuss a few potential future applications below, which have already shown promising results for simplified test cases.

The machine-learning community continues to discover powerful methods for processing and classifying complex data with inherent structure, such as trees and graphs. Complex data structures are prevalent at the LHC. The set of particles that make up a jet can be mapped to a tree structure. We have already discussed how RNNs can be used to identify jets that originate from beauty quarks, but this is just one of the many potential applications of RNNs, or of graph convolutional networks, to the study of jets[87].

Generative models, which learn the probability distribution of features directly, are capable of producing simulated data that closely approximate experimental data using tools such as generative adversarial networks[88] and variational auto-encoders[89,90]. A generative adversarial network uses one neural network, the 'generator', to generate potential data samples using random noise as input, while a second network, the 'adversary', penalizes the generator during training if the data that it generates can be distinguished from the training data. Although they are difficult to train, these networks can potentially generate large data samples much faster than can existing simulation tools, offering the possibility of providing the orders-of-magnitude-larger simulation samples that will be required by future experiments. Early work in this direction is encouraging[63,91,92], demonstrating that accurate simulations of a simplified calorimeter can be produced while achieving a marked decrease in the computational resources required.

The adversarial approach can also be applied to training classifiers with the ability to enforce invariance to latent parameters. This represents a new way of making classifiers robust against systematic uncertainties[93] and is a viable approach to avoid biasing a physical feature such as mass[65]. Several promising alternatives are also being investigated[94–97], some of which have been used for analysis at LHCb[98]. All of these approaches share the common theme of altering the training of the algorithms to reduce the potential bias learned. These are only a few of the machine-learning developments that are revolutionizing data interpretation in particle physics, greatly increasing the discovery potential of present and future experiments.

Received: 4 October 2017; Accepted: 19 June 2018;
Published online 1 August 2018.

1. Glaser, D. A. Some effects of ionizing radiation on the formation of bubbles in liquids. *Phys. Rev.* **87**, 665 (1952).
2. Evans, L. & Bryant, P. LHC machine. *J. Instrum.* **3**, S08001 (2008).
3. Alves, A. A. Jr et al. The LHCb detector at the LHC. *J. Instrum.* **3**, S08005 (2008).

© 2018 Springer Nature Limited. All rights reserved.

4. Aad, G. et al. The ATLAS experiment at the CERN Large Hadron Collider. *J. Instrum.* **3**, S08003 (2008).
5. Chatrchyan, S. et al. The CMS experiment at the CERN LHC. *J. Instrum.* **3**, S08004 (2008).
6. Bhat, P. Multivariate analysis methods in particle physics. *Annu. Rev. Nucl. Part. Sci.* **61**, 281–309 (2011).
7. Rosenblatt, F. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Spartan Books, Berlin, 1961).
8. Reed, R. & Marks, R. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks* (MIT Press, Cambridge, 1999).
9. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (Wadsworth International Group, Belmont, 1984).
10. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
11. The ALEPH Collaboration. Determination of $|V_{ub}|$ from the measurement of the inclusive charmless semileptonic branching ratio of b hadrons. *Eur. Phys. J. C* **6**, 555–574 (1999).
12. OPAL Collaboration. A measurement of the production of $D^{*\pm}$ mesons on the Z^0 resonance. *Z. Phys. C* **67**, 27–44 (1995).
13. Chiappetta, P., Colangelo, P., De Felice, P., Nardulli, G. & Pasquariello, G. Higgs search by neural networks at LHC. *Phys. Lett. B* **322**, 219–223 (1994).
14. Peterson, C., Rognvaldsson, T. & Lönnblad, L. JETNET 3.0—a versatile artificial neural network package. *Comput. Phys. Commun.* **81**, 185–220 (1994).
15. Buskulic, D. et al. Measurement of the tau polarisation at the Z resonance. *Z. Phys. C* **59**, 369–386 (1993).
16. Babbage, W. S. & Thompson, L. F. The use of neural networks in γ-π^0 discrimination. *Nucl. Instrum. Methods A* **330**, 482–486 (1993).
17. Lönnblad, L., Peterson, C. & Rognvaldsson, T. Pattern recognition in high energy physics with artificial neural networks — JETNET 2.0. *Comput. Phys. Commun.* **70**, 167–182 (1992).
18. Peterson, C. & Rögnvaldsson, T. S. An introduction to artificial neural networks. In *14th CERN School of Computing* (ed. Verkerk, C.) 113–170 (CERN, 1992).
19. Lönnblad, L., Peterson, C. & Rögnvaldsson, T. Using neural networks to identify jets. *Nucl. Phys. B* **349**, 675–702 (1991).
20. Lönnblad, L., Peterson, C. & Rögnvaldsson, T. Finding gluon jets with a neural trigger. *Phys. Rev. Lett.* **65**, 1321–1324 (1990).
21. Denby, B. Neural networks and cellular automata in experimental high-energy physics. *Comput. Phys. Commun.* **49**, 429–448 (1988).
22. Roe, B. P. et al. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nucl. Instrum. Methods A* **543**, 577–584 (2005).
23. Aad, G. et al. Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716**, 1–29 (2012).
24. Chatrchyan, S. et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716**, 30–61 (2012).
25. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
26. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
27. Vagata, P. & Wilfong, K. Scaling the Facebook data warehouse to 300 PB. *Facebook Code* https://code.fb.com/core-data/scaling-the-facebook-data-warehouse-to-300-pb/ (2014).
28. CMS Collaboration. Boosted decision trees in the level-1 muon endcap trigger at CMS. In *18th International Workshop on Advanced Computing and Analysis Techniques in Physics Research* 21–25 (CERN, 2017).
29. Aaij, R. et al. The LHCb trigger and its performance in 2011. *J. Instrum.* **8**, P04022 (2013).
30. Gligorov, V. V. & Williams, M. Efficient, reliable and fast high-level triggering using a bonsai boosted decision tree. *J. Instrum.* **8**, P02013 (2013).
This paper presents a boosted decision tree that identifies data in real time at LHCb and that has been used in more than 200 journal articles so far.
31. Likhomanenko, T. et al. LHCb topological trigger reoptimization. *J. Phys. Conf. Ser.* **664**, 082025 (2015).
32. The LHCb Collaboration. LHCb detector performance. *Int. J. Mod. Phys. A* **30**, 1530022 (2015).
33. Aaij, R. et al. Search for dark photons in 13 TeV *pp* collisions. *Phys. Rev. Lett.* **120**, 061801 (2018).
34. Hushchyn, M. et al. GRID storage optimization in transparent and user-friendly way for LHCb datasets. *J. Phys. Conf. Ser.* **898**, 062023 (2017).
35. Derkach, D. et al. LHCb trigger streams optimization. *J. Phys. Conf. Ser.* **898**, 062026 (2017).
36. Borisyak, M., Ratnikov, F., Derkach, D. & Ustyuzhanin, A. Towards automation of data quality system for CERN CMS experiment. *J. Phys. Conf. Ser.* **898**, 092041 (2017).
37. Kuznetsov, V. et al. Predicting dataset popularity for the CMS experiment. *J. Phys. Conf. Ser.* **762**, 012048 (2016).
38. Hushchyn, M., Charpentier, P. & Ustyuzhanin, A. Disk storage management for LHCb based on data popularity estimator. *J. Phys. Conf. Ser.* **664**, 042026 (2015).
39. Bonacorsi, D. et al. Monitoring data transfer latency in CMS computing operations. *J. Phys. Conf. Ser.* **664**, 032033 (2015).
40. CMS Collaboration. Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV. *J. Instrum.* **8**, P09009 (2013).
41. The ATLAS Collaboration. Evidence for the $H \rightarrow bb^-$ decay with the ATLAS detector. *J. High Energy Phys.* **12**, 24 (2017).
42. CMS Collaboration. *Evidence for the Decay of the Higgs Boson to Bottom Quarks.* Report No. CMS-PAS-HIG-16-044, https://cds.cern.ch/record/2278170 (CERN, 2017).
43. Aad, G. et al. Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector. *J. High Energy Phys.* **4**, 117 (2015).
44. Adam-Bourdarios, C. et al. The Higgs boson machine learning challenge. *J. Mach. Learn. Res. Worksh. Conf. Proc.* **42**, 19–55 (2014).
This work helped to popularize particle physics in the general machine-learning community and advertised recent advances in machine learning within the particle-physics community.
45. CMS Collaboration & LHCb Collaboration Observation of the rare $B_s^0 \rightarrow \mu^+\mu^-$ decay from the combined analysis of CMS and LHCb data. *Nature* **522**, 68–72 (2015).
46. Aaij, R. et al. Measurement of the $B_s^0 \rightarrow \mu^+\mu^-$ branching fraction and effective lifetime and search for $B^0 \rightarrow \mu^+\mu^-$ decays. *Phys. Rev. Lett.* **118**, 191801 (2017).
47. Yonghui, W. et al. Google's neural machine translation system: bridging the gap between human and machine translation. Preprint at https://arxiv.org/abs/1609.08144 (2016).
48. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
49. Baldi, P., Sadowski, P. & Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.* **5**, 4308 (2014).
This paper introduced deep learning to high-energy physics and explains the difference between shallow networks with high-level features and deep networks that find their own high-level features.
50. de Oliveira, L., Kagan, M., Mackey, L., Nachman, B. & Schwartzman, A. Jet-images — deep learning edition. *J. High Energy Phys.* **7**, 69 (2016).
This paper started investigations of deep-learning approaches to jets in quantum chromodynamics and includes a detailed discussion of CNNs and supporting exploration of network behaviour.
51. Racah, E. et al. Revealing fundamental physics from the Daya Bay Neutrino Experiment using deep neural networks. In *15th IEEE International Conference on Machine Learning and Applications* 892–897 (IEEE, 2016).
52. Aurisano, A. et al. A convolutional neural network neutrino event classifier. *J. Instrum.* **11**, P09001 (2016).
The paper presents the first CNN to be used for a physics analysis[70] and includes a detailed discussion of the method and comparison to more traditional neutrino-identification methods.
53. Sadowski, P, et al. Efficient antihydrogen detection in antimatter physics by deep learning. *J. Phys. Commun.* **1**, 025001 (2017).
54. Renner, J. Background rejection in NEXT using deep neural networks. *J. Instrum.* **12**, T01004 (2017).
55. Wielgosz, M., Skoczeń, A. & Mertik, M. Using LSTM recurrent neural networks for monitoring the LHC superconducting magnets. *Nucl. Instrum. Methods A* **867**, 40–50 (2017).
56. Edelen, A. L. et al. Neural networks for modeling and control of particle accelerators. *IEEE Trans. Nucl. Sci.* **63**, 878–897 (2016).
57. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989).
58. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: continual prediction with LSTM. *Neural Comput.* **12**, 2451–2471 (1999).
59. Cogan, J., Kagan, M., Strauss, E. & Schwartzman, A. Jet-images: computer vision inspired techniques for jet tagging. *J. High Energy Phys.* **2**, 118 (2015).
60. Baldi, P., Bauer, K., Eng, C., Sadowski, P. & Whiteson, D. Jet substructure classification in high-energy physics with deep neural networks. *Phys. Rev. D* **93**, 094034 (2016).
61. Barnard, J., Dawe, E. N., Dolan, M. J. & Rajcic, N. Parton shower uncertainties in jet substructure analyses with deep neural networks. *Phys. Rev. D* **95**, 014018 (2017).
62. Komiske, P. T., Metodiev, E. M. & Schwartz, M. D. Deep learning in color: towards automated quark/gluon jet discrimination. *J. High Energy Phys.* **1**, 110 (2017).
63. de Oliveira, L., Paganini, M. & Nachman, B. Learning particle physics by example: location-aware generative adversarial networks for physics synthesis. *Comput. Softw. Big Sci.* **1**, 4 (2017).
64. Kasieczka, G., Plehn, T., Russell, M. & Schell, T. Deep-learning top taggers or the end of QCD? *J. High Energy Phys.* **5**, 6 (2017).
65. Shimmin, C. et al. Decorrelated jet substructure tagging using adversarial neural networks. *Phys. Rev. D* **96**, 074034 (2017).
66. The ATLAS Collaboration. *Quark versus Gluon Jet Tagging using Jet Images with the ATLAS Detector.* Report No. ATL-PHYS-PUB-2017-017, https://cds.cern.ch/record/2275641 (CERN, 2017).
67. CMS Collaboration. *New Developments for Jet Substructure Reconstruction in CMS.* Report No. CMS-DP-2017-027, https://cds.cern.ch/record/2275226 (CERN, 2017).
68. NOvA Collaboration. *The NOvA Technical Design Report.* Report No. FERMILAB-DESIGN-2007-01 (FNAL, 2007)
69. Szegedy, C. et al. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition* 1–9 (IEEE, 2015).
70. Adamson, P. et al. Constraints on oscillation parameters from ν_e appearance and ν_μ disappearance in NOvA. *Phys. Rev. Lett.* **118**, 231801 (2017).
71. Acciarri, R. et al. Design and construction of the MicroBooNE detector. *J. Instrum.* **12**, P02017 (2017).
72. Adamson, P. et al. Search for active-sterile neutrino mixing using neutral-current interactions in NOvA. *Phys. Rev. D* **96**, 072006 (2017).

© 2018 Springer Nature Limited. All rights reserved.

73. Acciarri, R. et al. Convolutional neural networks applied to neutrino events in a liquid argon time projection chamber. *J. Instrum.* **12**, P03011 (2017).
74. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 061137 (2017).
75. ATLAS Collaboration. Performance of *b*-jet identification in the ATLAS experiment. *J. Instrum.* **11**, P04008 (2016).
76. CMS Collaboration. *CMS Phase 1 Heavy Flavour Identification Performance and Developments*. Report No. CERN-CMS-DP-2017-013, https://cds.cern.ch/record/2263802 (CERN, 2017).
77. ATLAS Collaboration. *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*. Report No. ATL-PHYS-PUB-2017-003, https://cds.cern.ch/record/2255226 (CERN, 2017).
78. ATLAS Collaboration. *Optimisation and Performance Studies of the ATLAS b-Tagging Algorithms for the 2017-18 LHC Run*. Report No. ATL-PHYS-PUB-2017-013, https://cds.cern.ch/record/2273281 (CERN, 2017).
79. CMS Collaboration. *Heavy Flavor Identification at CMS with Deep Neural Networks*. Report No. CMS-DP-2017-005, https://cds.cern.ch/record/2255736 (CERN, 2017).
80. Guest, D. et al. Jet flavor classification in high-energy physics with deep neural networks. *Phys. Rev. D* **94**, 112002 (2016).
81. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
82. ATLAS Collaboration. Electron efficiency measurements with the ATLAS detector using 2012 LHC proton–proton collision data. *Eur. Phys. J. C* **77**, 195 (2017).
83. The CMS Collaboration. Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV. *J. Instrum.* **10**, P06005 (2015).
84. Sachdev, K. *Muon Neutrino to Electron Neutrino Oscillation in NOvA*. PhD thesis, Univ. Minnesota (2015).
85. Chatrchyan, S. et al. Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons. *J. High Energy Phys.* **5**, 104 (2014).
86. van der Maaten, L. Accelerating t-SNE using tree- based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
87. Louppe, G., Cho, K., Becot, C. & Cranmer, K. QCD-aware recursive neural networks for jet physics. Preprint at https://arxiv.org/abs/1702.00748 (2017).
88. Goodfellow, I. J. et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**, 2672–2680 (2014).
89. Rezende, D.J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *J. Mach. Learn. Res. Worksh. Conf. Proc.* **32**, 1278–1286 (2014).
90. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2014).
91. Paganini, M., de Oliveira, L. & Nachman, B. Accelerating science with generative adversarial networks: an application to 3D particle showers in multilayer calorimeters. *Phys. Rev. Lett.* **120**, 042003 (2018).
92. Carminati, F. et al. Calorimetry with deep learning: particle classification, energy regression, and simulation for high-energy physics. In *NIPS Deep Learning for Physical Sciences Workshop* (NIPS, 2017).
93. Louppe, G., Kagan, M. & Cranmer, K. Learning to pivot with adversarial networks. *Adv. Neural Inf. Process. Syst.* **30**, 981–990 (2017).
This paper forms part of a collection of works that present a more nuanced loss function and, along with similar work for BDTs (see section 'Conclusions and outlook'), could lead to a new paradigm for training machine-learning models in high-energy physics.
94. Stevens, J. & Williams, M. uBoost: a boosting method for producing uniform selection efficiencies from multivariate classifiers. *J. Instrum.* **8**, P12013 (2013).
95. Rogozhnikov, A., Bukva, A., Gligorov, V. V., Ustyuzhanin, A. & Williams, M. New approaches for boosting to uniformity. *J. Instrum.* **10**, T03002 (2015).
96. Dery, L. M., Nachman, B., Rubbo, F. & Schwartzman, A. Weakly supervised classification in high energy physics. *J. High Energy Phys.* **5**, 145 (2017).
97. Baldi, P., Cranmer, K., Faucett, T., Sadowski, P. & Whiteson, D. Parameterized neural networks for high-energy physics. *Eur. Phys. J. C* **76**, 235 (2016).
98. Aaij, R. et al. Search for hidden-sector bosons in $B^0 \to K^{*0}\mu^+\mu^-$ decays. *Phys. Rev. Lett.* **115**, 161802 (2015).
99. The ATLAS collaboration. Search for the $b\bar{b}$ bb decay of the standard model Higgs boson in associated (W/Z)H production with the ATLAS detector. *J. High Energy Phys.* **1**, 69 (2015).
100. Chatrchyan, S. et al. Search for the standard model Higgs boson produced in association with a *W* or a *Z* boson and decaying to bottom quarks. *Phys. Rev. D* **89**, 012003 (2014).
101. CMS Collaboration. *2015 ECAL Detector Performance Plots*. Report No. CMS-DP-2015-057, https://cds.cern.ch/record/2114735 (CERN, 2015).

Reviewer information *Nature* thanks C. Backhouse, M. Pierini and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors contributed to writing this Review. A.R. and M.W. were the principal editors.

Competing interests The authors declare no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.R. or M.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0338-1

Genome-centric view of carbon processing in thawing permafrost

Ben J. Woodcroft[1,10], Caitlin M. Singleton[1,10], Joel A. Boyd[1], Paul N. Evans[1], Joanne B. Emerson[2,9], Ahmed A. F. Zayed[2], Robert D. Hoelzle[1], Timothy O. Lamberton[1], Carmody K. McCalley[3], Suzanne B. Hodgkins[4], Rachel M. Wilson[4], Samuel O. Purvine[5], Carrie D. Nicora[5], Changsheng Li[6], Steve Frolking[6], Jeffrey P. Chanton[4], Patrick M. Crill[7], Scott R. Saleska[8], Virginia I. Rich[2] & Gene W. Tyson[1]*

As global temperatures rise, large amounts of carbon sequestered in permafrost are becoming available for microbial degradation. Accurate prediction of carbon gas emissions from thawing permafrost is limited by our understanding of these microbial communities. Here we use metagenomic sequencing of 214 samples from a permafrost thaw gradient to recover 1,529 metagenome-assembled genomes, including many from phyla with poor genomic representation. These genomes reflect the diversity of this complex ecosystem, with genus-level representatives for more than sixty per cent of the community. Meta-omic analysis revealed key populations involved in the degradation of organic matter, including bacteria whose genomes encode a previously undescribed fungal pathway for xylose degradation. Microbial and geochemical data highlight lineages that correlate with the production of greenhouse gases and indicate novel syntrophic relationships. Our findings link changing biogeochemistry to specific microbial lineages involved in carbon processing, and provide key information for predicting the effects of climate change on permafrost systems.

Permafrost thaw induced by climate change is predicted to make up to 174 Pg of near-surface carbon (less than 3 m below the surface) available for microbial degradation by 2100[1]. Prediction of the magnitude of carbon loss as carbon dioxide (CO_2) or methane (CH_4) is hampered by our limited knowledge of microbial metabolism of organic matter in these environments. Genome-centric metagenomic analysis of microbial communities provides the necessary information to examine how specific lineages transform organic matter during permafrost thaw. However, these methods are challenged by the inherent complexity and spatial heterogeneity of near-surface soil communities that support diverse functional processes[2–4]. Previous metagenomic studies in permafrost-associated soils from Alaskan tundra[5,6] and a mineral soil permafrost[7] recovered a small number (14–33) of metagenome-assembled genomes (MAGs), which represent only a fraction of the species present in these systems. However, the ability to recover MAGs from complex microbial communities is continually improving in parallel with advances in sequencing technology and bioinformatic techniques[8,9].

Recovery and distribution of MAGs

The discontinuous permafrost at Stordalen Mire in northern Sweden is a model Arctic peatland ecosystem for studying thaw progression[10]. To gain an understanding of the microbial communities at this site and their associated carbon metabolism, soil samples were collected from the active layer (seasonally thawed) of three sites across a thaw gradient: an intact palsa (thawed to approximately 30 cm), a partially thawed bog (approximately 60 cm), and a fully thawed fen. Although bogs and fens exist in diverse landscapes, thaw-associated shifts in hydrology cause them to be a common feature of thawing northern peatland permafrost systems (see Methods). Triplicate soil cores and biogeochemical measurements were taken at each site from four active layer depths (near surface, mid, deep and extra-deep; 1–51 cm; Fig. 1, Supplementary Data 1) over three growing seasons. In total, 1.7 Tb of metagenomic sequence data were generated from 214 samples (2–165 Gb per sample), with supporting metatranscriptomic and metaproteomic data from a subset of these samples (Supplementary Data 1, 2). Metagenome assembly and differential coverage binning yielded 1,529 medium- to high-quality MAGs (more than 70% complete and less than 10% contaminated) from a diverse range of bacterial (1,434 genomes) and archaeal phyla (95 genomes; Extended Data Fig. 1a, Supplementary Data 3). The Stordalen Mire MAGs expand the number of genomes recovered from permafrost-associated soils by two orders of magnitude.

To resolve the taxonomic distribution of the Stordalen Mire MAGs, phylogenetic trees were inferred from concatenated sets of single-copy marker genes (120 bacterial or 122 archaeal genes). The recovered MAGs spanned 30 phyla, including Bacteria belonging to Actinobacteria (385 genomes), Acidobacteria (364), Proteobacteria (205) and Chloroflexi (66), and Archaea from the Euryarchaeota (85). The Stordalen genomes substantially expand representation of several common soil-dwelling lineages (Extended Data Fig. 1), such as the ubiquitous Acidobacteria, for which genomic representation was increased threefold. MAGs were also recovered from many poorly characterized phyla, including 47 genomes from the bacterial candidate phylum Dormibacteraeota (AD3), 53 from Eremiobacteraeota (WPS-2), six from FCPU426 and eight archaea from the Bathyarchaeota (Extended Data Fig. 1, Supplementary Data 10, 11). The Stordalen genomes broadly represent the major groups present in the system (Extended Data Fig. 2a, b) as well as many lineages previously detected in other permafrost-associated environments[6,11]. On the basis of the diversity of ribosomal protein sequences detected in the

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Queensland, Australia. [2]Department of Microbiology, The Ohio State University, Columbus, OH, USA. [3]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, USA. [4]Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA. [5]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA. [6]Earth Systems Research Center, Institute for the Study of Earth, Oceans and Space, University of New Hampshire, Durham, NH, USA. [7]Department of Geological Sciences and Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden. [8]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [9]Present address: Department of Plant Pathology, University of California, Davis, CA, USA. [10]These authors contributed equally: Ben J. Woodcroft, Caitlin M. Singleton. *e-mail: g.tyson@uq.edu.au

© 2018 Springer Nature Limited. All rights reserved.

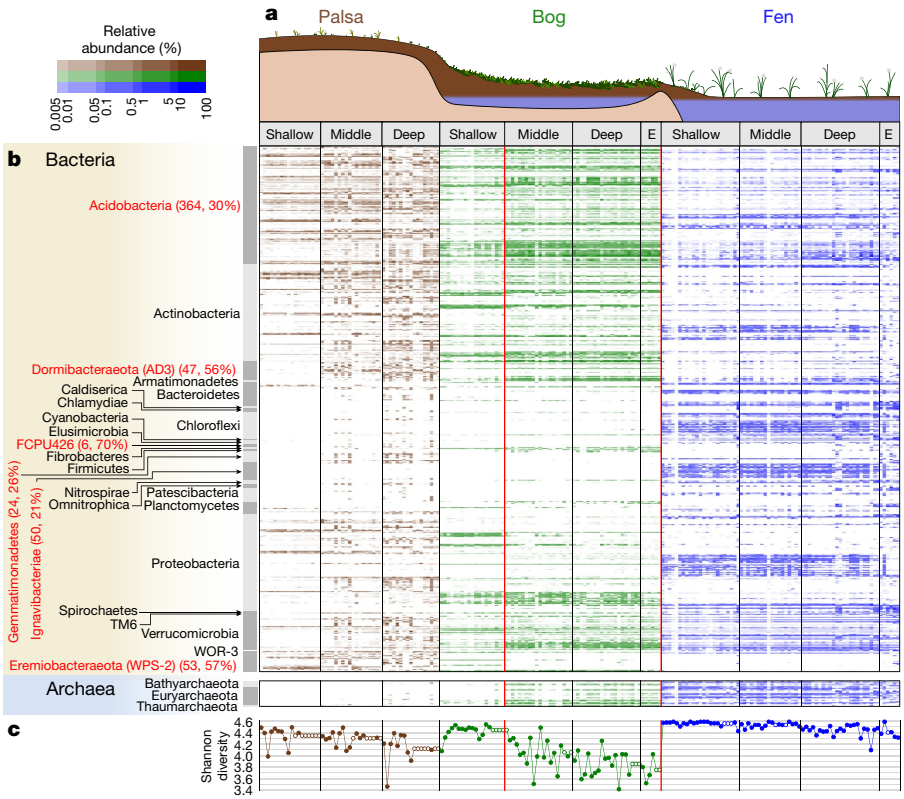


Fig. 1 | Genome-resolved view of the microbial communities at Stordalen Mire. a, Schematic of the permafrost thaw gradient. Permafrost is light brown, active layer is blue (saturated peat) and dark brown (non-saturated peat). **b**, Community profile derived from MAG abundances (rows) across the active layer metagenomes (columns) from palsa (brown), bog (green) and fen (blue) samples. Black lines divide sites and depth (E, extra-deep). Red lines separate samples taken above (left) and below (right) the water table. Numbers in parentheses show total MAGs recovered and phylogenetic gain of Stordalen MAGs compared to publicly available genomes for each phylum. Red text indicates previously poorly represented phyla. **c**, Shannon diversity of each sample (filled circles) or averages for the sample's thaw stage and depth (open circles).

metagenomes, we conservatively estimate that more than 24,000 strains inhabit Stordalen Mire (Supplementary Note 1). The Stordalen MAGs represent about 60% of microorganisms in the mire at the genus level (Supplementary Note 2), making this, to our knowledge, the most comprehensive recovery of genomes from a complex, natural soil environment to date.

Stordalen genomes were explicitly linked to the changing habitats by mapping their abundances across the thaw gradient (Fig. 1, Supplementary Data 4). Communities shifted substantially between sites (Extended Data Fig. 2c), with MAGs belonging to the Acidobacteria, Actinobacteria, Eremiobacteraeota, Alphaproteobacteria and Gammaproteobacteria predominant in the palsa and bog (5–41% of the community), whereas Deltaproteobacteria, Bacteroidetes, Chloroflexi and Ignavibacteriae were almost exclusively observed in the fen (8–14%; Extended Data Fig. 2e, f). The extra-deep bog samples had the lowest diversity (Shannon index 3.74 ± 0.24; Fig. 1), potentially owing to the ombrotrophic and anaerobic conditions in this environment, whereas the shallow samples from the minerotrophic fen were the most diverse (Shannon index 4.55 ± 0.05; Fig. 1). The fen also had 2.6 times more microbial cells per gram of soil relative to the palsa and bog (Extended Data Fig. 2d). In the bog, decreasing oxygen availability with depth is likely to drive changes in community structure, with Euryarchaeota and Dormibacteraeota increasing in relative abundance with depth, and Eremiobacteraeota decreasing. In the fen, the Planctomycetes, Omnitrophica and Spirochaetes increased in abundance with depth (Extended Data Fig. 2e, f). Consistent with the heterogeneity of soil environments[12,13], individual MAGs were typically found only at high abundance (over 1%) in a limited number of samples (fewer than 4). However, a small number of MAGs belonging to the Acidobacteria, Actinobacteria, Proteobacteria and Euryarchaeota were ubiquitous at specific depths within the palsa, bog or fen (Extended Data Fig. 3a, b; Extended Data Table 1). For several genera, closely related MAGs were abundant at different depths (Extended Data Fig. 3c), reflecting fine-scale adaptation to distinct niches in the soil column.

Polysaccharide degradation

Metabolic reconstruction of the MAGs, combined with 24 metatranscriptome and 16 metaproteome datasets (Extended Data Table 2), allowed examination of the key microorganisms, pathways and interactions responsible for organic matter degradation and the production of greenhouse gases at Stordalen Mire (Fig. 2; Supplementary Data 1). The first stage in degradation involves the breakdown of high-molecular-weight plant-derived polysaccharides, primarily cellulose and hemicellulose, which make up a large proportion of peatland carbon[14]. The ability to degrade these polysaccharides was a dominant feature of the Stordalen MAGs across all three thaw environments (Supplementary Data 5; Fig. 2 'MAG abundances' and 'distribution box plots'), with many encoding cellulases and xylanases (39% and 37%, respectively; average 3.8 and 2.6 copies per genome). This is consistent with gene-centric metagenomic studies of Arctic fens and tundras[15,16]; however, the genome-centric approach used here links these metabolic functions to specific populations.

Cellulase- and xylanase-encoding microorganisms, primarily belonging to the Acidobacteria (Fig. 2 'MAG abundances'), were most abundant in the palsa surface (68% and 59% of the recovered community, respectively), and decreased with depth (Fig. 2 'distribution box plots'). The surface bog had the lowest percentage of microorganisms encoding these genes (34% and 24%, respectively), probably owing to breakdown inhibition through the production of acids by *Sphagnum* moss[17]. The high relative abundance of cellulase- and xylanase-encoding Acidobacteria (61% and 75% of acidobacterial

© 2018 Springer Nature Limited. All rights reserved.

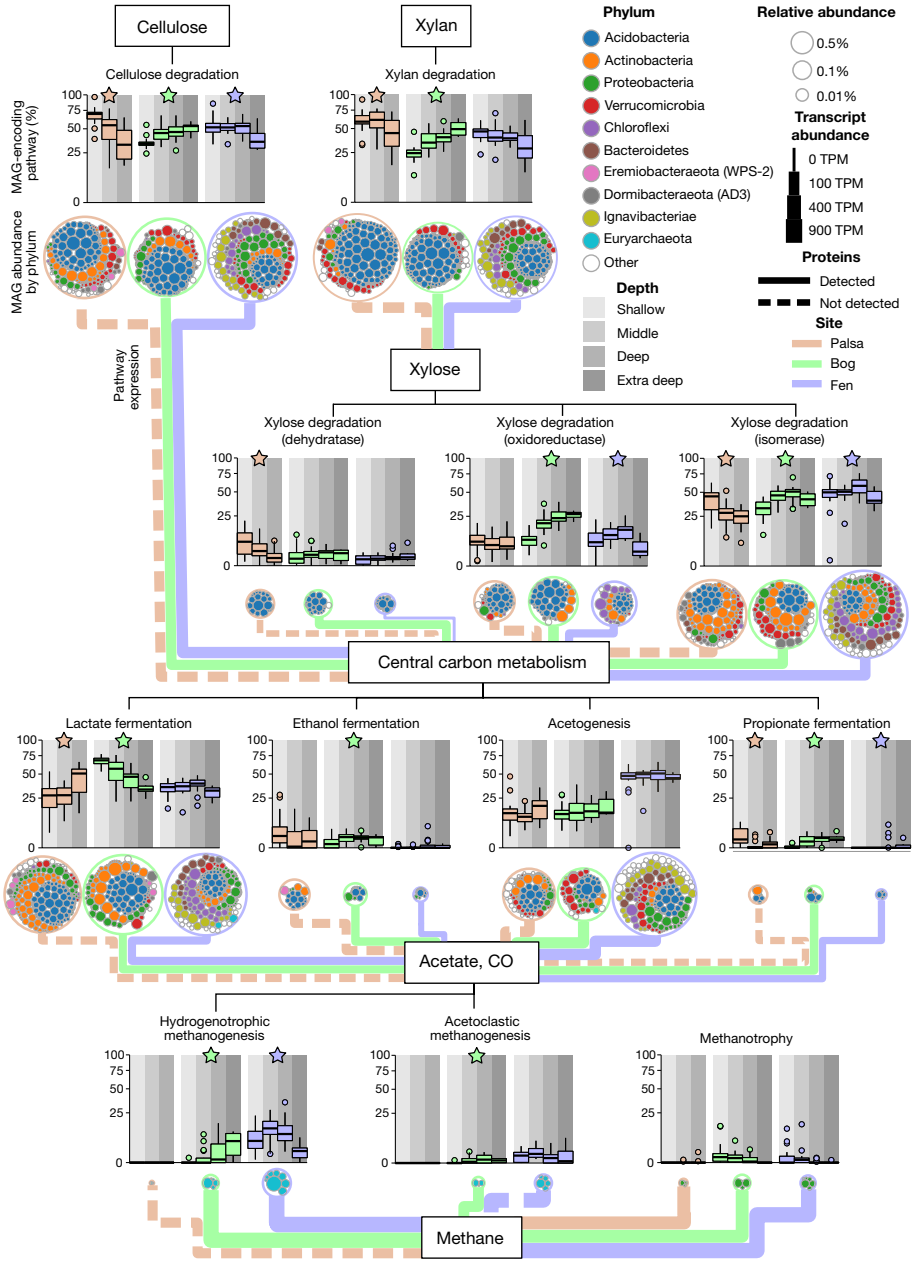


Fig. 2 | Carbon metabolism across the thaw gradient. Box plot headers and carbon compound boxes show degradation pathways. The large circles have outlines coloured by site, and contain smaller circles (MAG abundances, coloured by phylum) representing the different MAGs encoding genes for the pathways shown. Circle size indicates MAG average relative abundance. The distribution box plots are coloured by site and stratified by depth. Box plot *y* axes indicate the cumulative relative abundances of the MAGs encoding the pathway of interest at the sites. Line thickness connecting the intermediates represents the average relative transcript expression ('pathway expression') of pathway genes as transcripts per million reads mapped (TPM). Lines denote whether proteins were detected (solid) or not detected (dashed) in the metaproteomes. Relative abundances between sites were found to be significantly different for all pathways shown (see Methods). Coloured stars indicate relative abundances (ANOVA, $P < 0.05$) of pathways that are significantly different between depths.

genomes, respectively), strongly suggests that they are the primary degraders of large polysaccharides in the palsa and bog (Figs. 1, 2). Metatranscriptomic data confirmed these genomic inferences, with Acidobacteria producing the majority of cellulase and xylanase transcripts at these sites (Extended Data Fig. 4). Metaproteomic analysis revealed protein expression for 45 cellulases and 27 xylanases primarily belonging to Acidobacteria in the bog (Supplementary Data 2). A wider range of microorganisms are responsible for this functionality in the fen, including members of the Proteobacteria, Ignavibacteriae, Bacteroidetes, Verrucomicrobia, Chloroflexi, and Actinobacteria. However, the metatranscriptomic data indicate that the Proteobacteria, Ignavibacteriae and Bacteroidetes have the highest expression of these genes in the fen, although only a limited number of proteins were detected. Notably, most putative cellulose hydrolysers also encode a xylanase (59% of genomes), with the exception of actinobacterial hydrolysers, which typically encode only cellulases (87%). However, unlike findings in other Arctic systems[18], both metatranscriptomic and metaproteomic data show that the actinobacterial cellulases are not highly expressed, indicating that these microorganisms play a minor role in polysaccharide degradation at Stordalen Mire. The high abundance of hydrolysers in the palsa suggests that the microbial community contributes to physical compaction through decomposition of surface organic matter, as evidenced by increases in bulk density with depth (Supplementary Note 3), and that this contribution is likely to augment thawing of underlying permafrost as the primary driver of subsidence.

© 2018 Springer Nature Limited. All rights reserved.

Breakdown of polysaccharides into simple sugars is the primary source of energy and carbon for the microbial community[16]. β-Glucosidases for disaccharide degradation were encoded by the majority of microorganisms in all sites (75%, 84% and 66% of palsa, bog and fen communities, respectively), with transcript expression primarily by Acidobacteria in the palsa and bog, and Bacteroidetes in the fen (Extended Data Fig. 4). The metaproteomes supported the high expression of β-glucosidase proteins by Acidobacteria in the bog (198 out of a total of 216 detected proteins). Degradation pathways for the monosaccharides glucose, galactose and xylose were also prevalent in the MAGs (Supplementary Note 4, Extended Data Fig. 5). Of the 237 microorganisms potentially capable of xylan degradation, 108 appear to be involved in xylose degradation using the canonical bacterial isomerase pathway[19] (Extended Data Fig. 6a). These genomes were common in the surface palsa (40% of the microbial community), deep bog (49%) and throughout the fen (51%), similar to the distribution of microorganisms capable of degrading the precursor xylan. Transcription of this pathway in the fen was highest by Bacteroidetes and Ignavibacteriae (Fig. 2 'pathway expression'; Extended Data Fig. 6). Members of the Acidobacteria, Actinobacteria and Verrucomicrobia showed highest transcription of this pathway in the bog, whereas expression was limited to Actinobacteria in the palsa (Extended Data Fig. 6c–h). Notably, 50 actinobacterial MAGs encoded genes necessary for xylose degradation, despite 44 being unable to degrade the precursor xylan, indicating that they are reliant on the activity of xylan hydrolysers.

Only a small fraction of the dominant acidobacterial xylan hydrolysers encode the necessary genes for the canonical isomerase pathway for xylose degradation (30 out of 111 genomes). Few acidobacterial genomes (23) encoded the alternative xylonate dehydratase pathway, indicating that they might degrade xylose through a membrane-bound glucose dehydrogenase, as previously observed only in *Gluconobacter oxydans*[20] (Supplementary Note 5). A closer inspection of acidobacterial xylan-hydrolysing MAGs revealed an oxidoreductase pathway for the conversion of xylose into xylulose, previously identified only in fungi[21–23] (37 genomes; Supplementary Note 5). MAGs belonging to the Actinobacteria and Chloroflexi also encoded this pathway, together comprising 13% of the community across the thaw gradient. Acidobacterial and actinobacterial genes for the oxidoreductase pathway were expressed in metatranscriptomes from across the mire, and were more highly expressed than the canonical isomerase pathway in both the palsa and bog (Fig. 2 'pathway expression', Extended Data Fig. 6c–h). Nine MAGs expressed proteins for this pathway, primarily in the bog, confirming that this novel pathway is in use and is likely to account for a substantial fraction of xylose degradation at the mire (Supplementary Data 2). The detection and expression of several distinct pathways for xylose degradation, often occurring in the same genome (Extended Data Fig. 6b), reveals that one or multiple pathways may be active under specific environmental conditions (Supplementary Note 5).

Fermentation

In the anaerobic layers of the peat column, where inorganic terminal electron acceptors (TEAs) are rare[24,25], fermentation and acetogenesis are essential pathways for the further degradation of monosaccharides, and supply the substrates for methanogenesis. Fermentation produces low-molecular-weight alcohols and organic acids such as ethanol, propionate, acetate and lactate, as well as hydrogen and carbon dioxide[24,26]. In the palsa and bog, lactate fermentation is a common metabolism encoded by actinobacterial and acidobacterial MAGs (Fig. 2 'MAG abundances'), which are particularly abundant in the bog surface (36% and 16% of the community, respectively), but decrease with depth (Fig. 2 'distribution box plots'). Transcript expression of this pathway, while low across all sites, appears to be mostly limited to these lineages in the bog (Extended Data Fig. 7). Conversely, populations belonging to the Chloroflexi, Ignavibacteriae, Bacteriodetes, and Proteobacteria appear to be the primary lactate metabolizers in the fen (9%, 7%, 5% and 5%, respectively), with Ignavibacteriae the most transcriptionally active (Extended Data Fig. 7). A small fraction of Stordalen genomes are capable of ethanol and propionate fermentation (Fig. 2 'MAG abundances'), and expression of these pathways is low and mainly limited to the palsa and bog (Extended Data Fig. 7). The abundance of these microorganisms in the palsa suggests that they are potentially important fermenters during the early stages of thaw (Fig. 2 'distribution box plots'). Acetogens were most abundant in the fen across all depths, which suggests increased acetate production and is consistent with a preference for pH-neutral environments[24,25] (Fig. 2 'distribution box plots'). Fen acetogens belong to the Ignavibacteriae, Bacteroidetes and Verrucomicrobia, whereas in the palsa and bog this metabolism was limited to a small number of Acidobacteria, Actinobacteria and Verrucomicrobia (Fig. 2 'MAG abundances'). These distributions were also observed in the metatranscriptomes (Extended Data Fig. 7), with Bacteroidetes, Ignavibacteriae and Proteobacteria contributing to the slightly higher expression of acetogenesis transcripts in the fen, compared to expression by Acidobacteria and Verrucomicrobia in the bog (Extended Data Fig. 7). While acetate can be oxidized using available inorganic TEAs[27] (for example, sulphate or nitrate), these are at very low concentrations in Stordalen Mire (Supplementary Data 6). The unexpectedly high ratio of CO_2 to CH_4 produced at the site[28] (16:1 in the bog and 7:1 in the fen; Extended Data Fig. 8a) may signal the oxidation of fermentation products including acetate using organic TEAs such as humic substances[27,29,30].

Methane metabolism

Methanogenesis is the final step in anaerobic carbon transformation and is of critical concern in thawing permafrost peatland systems where CH_4 release is increasing rapidly[31]. Of the 95 archaeal genomes recovered (Extended Data Fig. 1), 76 were identified as hydrogenotrophic methanogens (H_2- and CO_2-utilizing), which alongside high transcript and protein expression (Extended Data Fig. 8d, e) suggests that this is the dominant form of CH_4 generation at the mire. Hydrogenotrophic methanogens increased in abundance, diversity and activity as thaw progressed from bog to fen (Fig. 2 'MAG abundances'), consistent with the increase in CH_4 flux[10], with mid- and deep-fen samples having the highest relative abundance of these methanogens (Fig. 2 'distribution box plots'). Only six low-abundance acetoclastic methanogens were recovered, primarily from the fen, where acetogenesis was also prevalent (Fig. 2 'MAG abundances'). In addition, two H_2-dependent methylotrophic methanogens from the order *Methanomassiliicoccales* were recovered, but were present at very low abundance (0.1% in the fen) with low transcriptional activity, making it unlikely that they contribute substantially to CH_4 production at the mire (Extended Data Fig. 8d). Methanotrophs from the Alphaproteobacteria and Gammaproteobacteria were identified across the thaw gradient. High abundances in the bog suggest that methanotrophs may oxidize substantial proportions of CH_4, limiting emissions to the atmosphere (Supplementary Note 6).

Microbial and geochemical interactions

The activity of methanogens and methanotrophs alters the $^{13}C/^{12}C$ isotopic ratio of CH_4 dissolved in the porewater[32]. A previous 16S rRNA gene amplicon survey at Stordalen Mire revealed that the abundance of *Candidatus* (*Ca.*) 'Methanoflorens stordalenmirensis' was the best single predictor of carbon isotopic fractionation during CH_4 production at the bog in 2011[33,34]. The recovery of 51 additional genomes here greatly expands the representation of the order *Ca.* 'Methanoflorentales', and revealed the presence of two habitat-specific clades derived from the bog and fen, respectively (80–85% average amino acid identity (AAI); *Ca.* 'M. stordalenmirensis' from the bog and *Ca.* 'Methanoflorens crillii' from the fen). The 16S rRNA gene-based correlation of *Ca.* 'M. stordalenmirensis' to the isotopic signature of CH_4 was confirmed by the metagenomic data for both 2011 and 2012, with the relative abundance of 19 *Ca.* 'M. stordalenmirensis' MAGs in the bog explaining more variation than bulk environmental variables (2011 $R^2 = 0.43$, $P = 6 \times 10^{-4}$; 2012 $R^2 = 0.48$, $P = 2 \times 10^{-4}$; Extended Data Fig. 8c, e). Notably, the relative abundance of a previously uncharacterized

© 2018 Springer Nature Limited. All rights reserved.

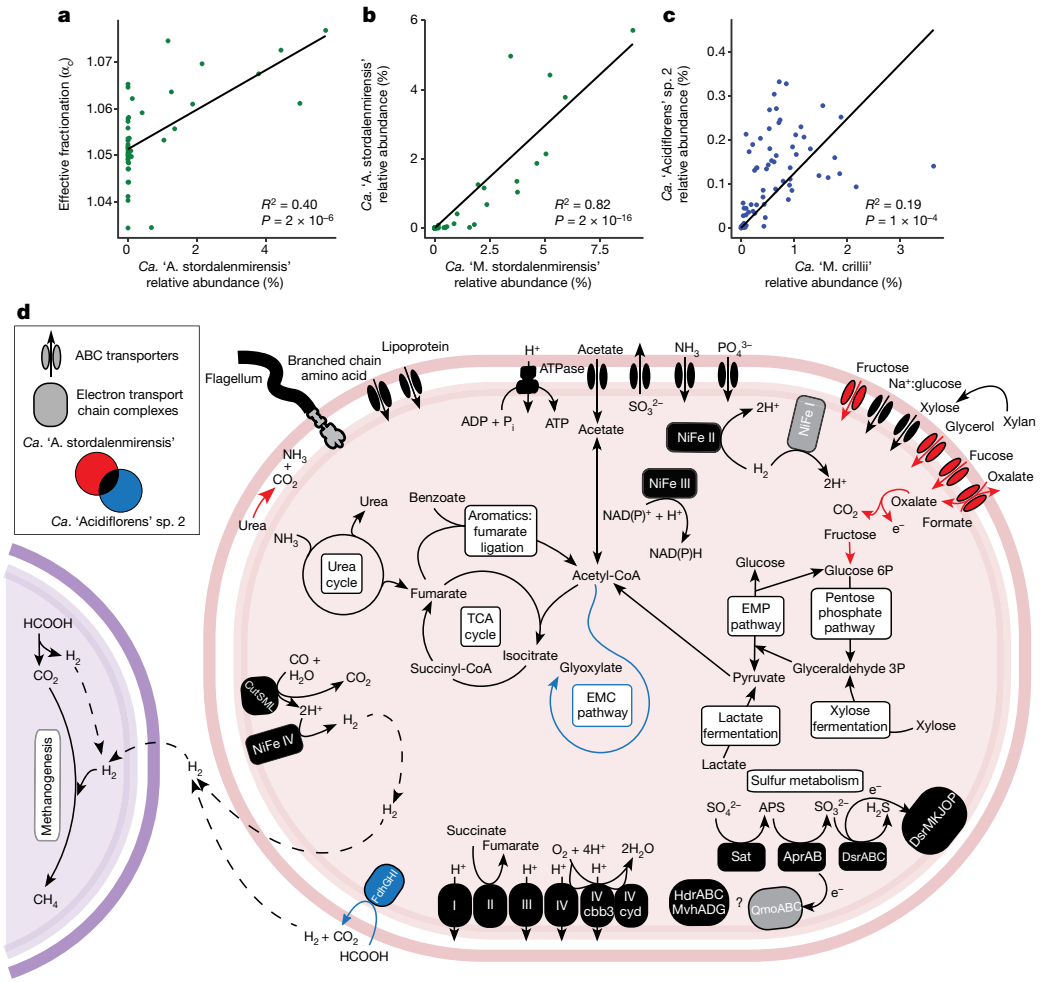


Fig. 3 | *Ca.* 'Acidiflorens' geochemical correlations and metabolic reconstruction. a, Correlation of relative abundances of *Ca.* 'Acidiflorens stordalenmirensis' with $\delta^{13}C$ of porewater CH_4 in bog sites. **b**, Correlation of the relative abundances of *Ca.* 'Methanoflorens stordalenmirensis' and *Ca.* 'Acidiflorens stordalenmirensis' in bog samples. **c**, Correlation of the relative abundances of *Ca.* 'Methanoflorens crillii' and *Ca.* 'Acidiflorens' sp. 2 in fen samples. **d**, Metabolic reconstruction of *Ca.* 'Acidiflorens'. Differential gene presence for the two lineages is indicated by colour-coding (Venn diagram). Functional units shown in grey are absent. Dotted lines indicate enzyme-independent hydrogen movement. Purple cell cartoon shows *Ca.* 'Methanoflorens' spp. consuming hydrogen produced by *Ca.* 'Acidiflorens' spp.

acidobacterial population, *Ca.* 'Acidiflorens stordalenmirensis', was significantly correlated with the isotopic composition of CH_4 ($R^2 = 0.40$, $P = 2 \times 10^{-6}$), and even more strongly correlated with the relative abundance of *Ca.* 'M. stordalenmirensis' ($R^2 = 0.82$, $P < 2 \times 10^{-16}$ in bog sites; Fig. 3, Supplementary Note 7).

A detailed metabolic analysis of *Ca.* 'A. stordalenmirensis' and the 49 other genomes belonging to the genus *Ca.* 'Acidiflorens' revealed metabolic capabilities that suggest that members of this lineage are facultative syntrophs (Extended Data Fig. 1b). Members of *Ca.* 'Acidiflorens' contain genes for the fermentation of a wide range of substrates including xylan, fatty acids, oxalate and fructose, and encode numerous hydrogenases, indicating the potential for H_2 production and consumption (Fig. 3d, Supplementary Note 8). We hypothesize that the correlation between *Ca.* 'A. stordalenmirensis' and *Ca.* 'M. stordalenmirensis' and the CH_4 isotopic composition indicates that these lineages are in a syntrophic relationship based on inter-species hydrogen transfer. Hydrogen consumption by *Ca.* 'M. stordalenmirensis' is likely to lower the hydrogen partial pressure and thereby make fermentation more thermodynamically favourable for *Ca.* 'A. stordalenmirensis'[35,36] (Fig. 3d). This syntrophy is also observed in the fen sites between closely related populations, as the relative abundances of a *Ca.* 'Methanoflorens' species (*Ca.* 'M. crillii'; see above) and a second *Ca.* 'Acidiflorens' species (*Ca.* 'Acidiflorens' sp. 2), both of which are detected only in the fen, were also correlated ($R^2 = 0.19$, $P = 1 \times 10^{-4}$; Fig. 3c). The species-level resolution afforded by the genome-centric metagenomic approach allowed identification of potential interactions between microorganisms and biogeochemistry that would have been missed using traditional gene amplicon surveys.

Another key but poorly constrained biogeochemical parameter in global CH_4 models[37] is the percentage of carbon mineralized to CO_2 versus CH_4. We directly examined the relationship of microbial lineages in the bog with the porewater CH_4:CO_2 ratio and identified a significant positive correlation with a genus within the candidate phylum Dormibacteraeota[38], named here *Ca.* 'Changshengia' ($R^2 = 0.19$, $P = 0.001$; Extended Data Fig. 9a, c). *Ca.* 'Changshengia' was found to oxidize glycerol, an important cryoprotectant in this Arctic environment[39,40], and its derivatives glycerol 3-phosphate and dihydroxyacetone. On the basis of the transcript and protein expression of genes for glycerol oxidation (Extended Data Fig. 9b, Supplementary Note 9), it is possible that *Ca.* 'Changshengia' ferments glycerol leading to the production[41] of H_2, which is transferred to methanogens, increasing the CH_4 to CO_2 ratio in the porewater.

Conclusion

Here, genome-centric metagenomic analysis of a permafrost thaw gradient allowed the recovery of 1,529 MAGs, substantially increasing the number of genomes sequenced from permafrost-associated environments. Analysis of these genomes and their abundances and expression enabled us to identify correlations between specific microbial populations and biogeochemistry, and revealed

© 2018 Springer Nature Limited. All rights reserved.

key populations that drive the mineralization of organic matter from plant-derived polysaccharides through to simple sugars, and the greenhouse gases CO_2 and CH_4. Future efforts that combine genome-centric meta-omic data with metabolomics and biogeochemical data will further improve our understanding of large-scale complex global processes, and inform Earth-system models for accurate predictions of climate-induced change.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0338-1.

Received: 15 June 2017; Accepted: 5 June 2018;
Published online 16 July 2018.

1. Schuur, E. A. et al. Climate change and the permafrost carbon feedback. *Nature* **520**, 171–179 (2015).
2. Roesch, L. F. et al. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* **1**, 283–290 (2007).
3. Delgado-Baquerizo, M. et al. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nat. Commun.* **7**, 10541 (2016).
4. Howe, A. C. et al. Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl Acad. Sci. USA* **111**, 4904–4909 (2014).
5. Johnston, E. R. et al. Metagenomics reveals pervasive bacterial populations and reduced community diversity across the Alaska tundra ecosystem. *Front. Microbiol.* **7**, 579 (2016).
6. Taş, N. et al. Landscape topography structures the soil microbiome in arctic polygonal tundra. *Nat. Commun.* **9**, 777 (2018).
7. Hultman, J. et al. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **521**, 208–212 (2015).
8. Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
9. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
10. Johansson, T. et al. Decadal vegetation changes in a northern peatland, greenhouse gas fluxes and net radiative forcing. *Glob. Change Biol.* **12**, 2352–2369 (2006).
11. Jansson, J. K. & Taş, N. The microbial ecology of permafrost. *Nat. Rev. Microbiol.* **12**, 414–425 (2014).
12. Whalen, S. C. Biogeochemistry of methane exchange between natural wetlands and the atmosphere. *Environ. Eng. Sci.* **22**, 73–94 (2005).
13. Rosenzweig, M. L. *Species Diversity in Space and Time* (Cambridge Univ. Press, Cambridge, 1995).
14. Kremer, C., Pettolino, F., Bacic, A. & Drinnan, A. Distribution of cell wall components in *Sphagnum* hyaline cells and in liverwort and hornwort elaters. *Planta* **219**, 1023–1035 (2004).
15. Tveit, A., Schwacke, R., Svenning, M. M. & Urich, T. Organic carbon transformations in high-Arctic peat soils: key functions and microorganisms. *ISME J.* **7**, 299–311 (2013).
16. Ivanova, A. A., Wegner, C. E., Kim, Y., Liesack, W. & Dedysh, S. N. Identification of microbial populations driving biopolymer degradation in acidic peatlands by metatranscriptomic analysis. *Mol. Ecol.* **25**, 4818–4835 (2016).
17. Pankratov, T. A., Ivanova, A. O., Dedysh, S. N. & Liesack, W. Bacterial populations and environmental factors controlling cellulose degradation in an acidic *Sphagnum* peat. *Environ. Microbiol.* **13**, 1800–1814 (2011).
18. Tveit, A., Urich, T. & Svenning, M. M. Metatranscriptomic analysis of Arctic peat soil microbiota. *Appl. Environ. Microbiol.* **80**, 5761–5772 (2014).
19. Jeffries, T. W. in *Pentoses and Lignin* 1–32 (Springer, Heidelberg, 1983).
20. Zhang, M. et al. Genetic analysis of D-xylose metabolism pathways in *Gluconobacter oxydans* 621H. *J. Ind. Microbiol. Biotechnol.* **40**, 379–388 (2013).
21. Kricka, W., Fitzpatrick, J. & Bond, U. Metabolic engineering of yeasts by heterologous enzyme production for degradation of cellulose and hemicellulose from biomass: a perspective. *Front. Microbiol.* **5**, 174 (2014).
22. Kuhn, A., van Zyl, C., van Tonder, A. & Prior, B. A. Purification and partial characterization of an aldo-keto reductase from *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **61**, 1580–1585 (1995).
23. Sarthy, A. V., Schopp, C. & Idler, K. B. Cloning and sequence determination of the gene encoding sorbitol dehydrogenase from *Saccharomyces cerevisiae*. *Gene* **140**, 121–126 (1994).
24. Ye, R. et al. pH controls over anaerobic carbon mineralization, the efficiency of methane production, and methanogenic pathways in peatlands across an ombrotrophic–minerotrophic gradient. *Soil Biol. Biochem.* **54**, 36–47 (2012).
25. Horn, M. A., Matthies, C., Küsel, K., Schramm, A. & Drake, H. L. Hydrogenotrophic methanogenesis by moderately acid-tolerant methanogens of a methane-emitting acidic peat. *Appl. Environ. Microbiol.* **69**, 74–83 (2003).
26. Conrad, R. Contribution of hydrogen to methane production and control of hydrogen concentrations in methanogenic soils and sediments. *FEMS Microbiol. Ecol.* **28**, 193–202 (1999).
27. Keller, J. K. & Takagi, K. K. Solid-phase organic matter reduction regulates anaerobic decomposition in bog soil. *Ecosphere* **4**, 54 (2013).
28. Hodgkins, S. B. et al. Changes in peat chemistry associated with permafrost thaw increase greenhouse gas production. *Proc. Natl Acad. Sci. USA* **111**, 5819–5824 (2014).
29. Lipson, D. A., Jha, M., Raab, T. K. & Oechel, W. C. Reduction of iron (III) and humic substances plays a major role in anaerobic respiration in an Arctic peat soil. *J. Geophys. Res. Biogeosci.* **115**, G00I06 (2010).
30. Klüpfel, L., Piepenbrock, A., Kappler, A. & Sander, M. Humic substances as fully regenerable electron acceptors in recurrently anoxic environments. *Nat. Geosci.* **7**, 195–200 (2014).
31. Christensen, T. R. et al. Thawing sub-arctic permafrost: effects on vegetation and methane emissions. *Geophys. Res. Lett.* **31**, L04501 (2004).
32. Whiticar, M. J. Carbon and hydrogen isotope systematics of bacterial formation and oxidation of methane. *Chem. Geol.* **161**, 291–314 (1999).
33. McCalley, C. K. et al. Methane dynamics regulated by microbial community response to permafrost thaw. *Nature* **514**, 478–481 (2014).
34. Mondav, R. et al. Discovery of a novel methanogen prevalent in thawing permafrost. *Nat. Commun.* **5**, 3212 (2014).
35. Stams, A. J. & Plugge, C. M. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nat. Rev. Microbiol.* **7**, 568–577 (2009).
36. Ishii, S., Kosaka, T., Hori, K., Hotta, Y. & Watanabe, K. Coaggregation facilitates interspecies hydrogen transfer between *Pelotomaculum thermopropionicum* and *Methanothermobacter thermautotrophicus*. *Appl. Environ. Microbiol.* **71**, 7838–7845 (2005).
37. Wania, R. et al. Present state of global wetland extent and wetland methane modelling: methodology of a model inter-comparison project (WETCHIMP). *Geosci. Model Dev.* **6**, 617–641 (2013).
38. Ji, M. et al. Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* **552**, 400–403 (2017).
39. Welsh, D. T. Ecological significance of compatible solute accumulation by micro-organisms: from single cells to global climate. *FEMS Microbiol. Rev.* **24**, 263–290 (2000).
40. Rodrigues, D. F. et al. Architecture of thermal adaptation in an *Exiguobacterium sibiricum* strain isolated from 3 million year old permafrost: a genome and transcriptome approach. *BMC Genomics* **9**, 547 (2008).
41. Maru, B., Bielen, A., Constanti, M., Medina, F. & Kengen, S. Glycerol fermentation to hydrogen by *Thermotoga maritima*: proposed pathway and bioenergetic considerations. *Int. J. Hydrogen Energy* **38**, 5563–5572 (2013).

Acknowledgements This study was funded by the Genomic Science Program of the United States Department of Energy (DOE) Office of Biological and Environmental Research (BER), grants DE-SC0004632, DE-SC0010580 and DE-SC0016440. B.J.W. and P.N.E. are supported by Australian Research Council Discovery Early Career Research Awards #DE160100248 and #DE170100428, respectively. C.M.S. and J.A.B. are supported by Australian Government Research Training Program (RTP) Scholarships, and G.W.T. is supported by Australian Research Council Future Fellowship FT170100070. A portion of the research was performed using Environmental Molecular Sciences Laboratory (EMSL), a DOE Office of Science User Facility, and a portion was performed under the Facilities Integrating Collaborations for User Science (FICUS) initiative with resources at both the DOE Joint Genome Institute and EMSL. Both facilities are sponsored by the Office of BER and operated under contracts DE-AC02-05CH11231 (JGI) and DE-AC05-76RL01830 (EMSL). We thank the IsoGenie 1 and 2 Project Teams and the 2010–2012 field teams for sample collection, particularly T. Logan, as well as the Abisko Scientific Research Station for sampling infrastructure and support. We thank P. Hugenholtz, D. Parks, S. Robbins, B. Kemish, M. Chuvochina, S. Low and M. Butler for helpful discussion and infrastructure support.

Reviewer information *Nature* thanks S. Allison, J. Jansson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.L., S.F., J.P.C., P.M.C., S.R.S., V.I.R. and G.W.T. designed the overall study of microbial and biogeochemical dynamics of permafrost thaw and procured funding. V.I.R. coordinated sampling efforts, and S.B.H., J.P.C. and G.W.T. collected samples. B.J.W., C.M.S., J.A.B., V.I.R. and G.W.T. designed experiments around specific microbial hypotheses. J.B.E., A.A.F.Z., S.O.P., and C.D.N performed the protein extractions and analyses. C.K.M., S.B.H. and R.M.W. performed geochemical analyses. B.J.W., C.M.S., J.A.B., P.N.E, R.D.H. and T.O.L. carried out microbial experiments and integration of microbial and geochemical data. B.J.W., C.M.S. and G.W.T. wrote the manuscript with contributions from J.A.B., P.N.E. and R.D.H. All authors except C.L. edited, reviewed and approved the final manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0338-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0338-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to G.W.T.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Study site. As described previously[34], Stordalen Mire is a peatland in northern Sweden (68 22′ N, 19 03′ E), 10 km southeast of Abisko. The three sub-habitats of the study site are common to northern wetlands, and together cover ~98% of Stordalen Mire's non-lake surface[10]. These proceed from well-drained palsas underlain by permafrost, and dominated by ericaceous and woody plants, to intermediate-thaw bogs with variable water table depth, dominated by *Sphagnum* spp. mosses, to fully thawed and inundated fens dominated by sedges such as *Eriophorum angustifolium*. A thaw-associated shift in these habitats was documented between 1970 and 2000, as palsa collapsed and bogs and fens expanded by 3% and 54%, respectively[10]. These habitats exist in an intermingled mosaic, as is common in discontinuous permafrost zones, and the specific palsa, bog and fen that were sampled in this study are directly adjacent such that all cores were collected within a 120 m radius.

This formation of wetlands after permafrost thaw is a widespread characteristic of peatlands affected by permafrost loss[42–45]. As frozen ground thaws it collapses, forming bogs and fens. Where this subsidence increases hydrologic connectivity, as at Stordalen, it can create a progression from ombrotrophic bogs to minerotrophic fens. A similar successional shift from bogs dominated by *Sphagnum* spp. to tall sedge fens has been observed in other northern peatlands[42,43,46,47]. The uncertainty surrounding the extent and characteristics of wetland formation from permafrost thaw is a critical limitation to modelling and understanding carbon–climate feedbacks[48,49]. Improved characterization of post-thaw microbial communities and carbon transformation processes, as advanced in this study, can directly address this uncertainty.

Geochemistry. Across the thaw chronosequence, porewater CH_4 and CO_2 measurements and their ^{13}C isotopic composition were sampled as described previously[28]. The $\delta^{13}C$-CH_4 is affected by the $\delta^{13}C$-CO_2, because of the use or production of CO_2 during CH_4 generation, so the isotopic fractionation factor is used to report the isotopic separation of CH_4 and carbon dioxide[50]. The α_C value reports the effective fractionation of C in CH_4, as the $\delta^{13}C$-CH_4 relative to source material represented by $\delta^{13}C$-CO_2. The effective fractionation factor of carbon in the porewater CH_4 relative to CO_2 (α_C) was calculated as described previously[33,50].

$$\alpha_c = \frac{\delta^{13}\mathrm{C} - \mathrm{CO}_2 + 1{,}000}{\delta^{13}\mathrm{C} - \mathrm{CH}_4 + 1{,}000}$$

DNA extraction and metagenome sequencing. DNA extractions were undertaken as described previously[34], with additional extractions from samples taken in 2012. Metagenome sequencing was performed for 2011 and 2012 using 100 ng of the DNA in TruSeq Nano (Illumina) library preparation. For low concentration DNA samples, libraries were created using 1 ng of DNA with the Nextera XT DNA Sample Preparation Kit (Illumina). 2012 libraries were sequenced on 1/12th of an Illumina HiSeq2000 lane producing 100 bp paired-end reads, although some 2012 and 2011 samples were selected for deeper sequencing. Libraries from 2011 were sequenced 1/24th of an Illumina NextSeq, producing 150 bp paired end reads. See Supplementary Data 1 for details of sequencing depth per sample.

Quantitative real-time PCR. A quantitative polymerase chain reaction (qPCR) analysis was performed on selected samples to quantify microbial load. After pre-diluting 1/100, PCR was set up using 5 μl of 2× SYBR Green/AmpliTaq Gold DNA Polymerase mix (Life Technologies, Applied Biosystems), 4 μl of microbial template DNA and 1 μl of primer mix. The 16S 1406F/1525R primer set (0.4 μM) was designed to amplify bacterial and archaeal 16S rRNA genes: F - GYACWCACCGCCCGT and R - AAGGAGGTGWTCCARCC. The rpsL F/R primer set (0.2 μM), used for inhibition control, amplifies *Escherichia coli* DH10B only: F - GTAAAGTATGCCGTGTTCGT and R - AGCCTGCTTACGGTCTTTA. Three dilutions, 1/100, 1/500 and 1/1,000 (microbial template DNA, 16S 1406F/1525R primer set) as well as an inhibition control (*E. coli* DH10B genomic DNA, rpsL primer set), were run in triplicate for each sample. The PCR was run on the ViiA7 platform (Applied Biosystems) including a cycle of 10 min at 95 °C (AmpliTaq activation) and 40 cycles of (15 s at 95 °C followed by 20 s at 55 °C and 30 s at 72 °C). A melt curve was produced by running a cycle of 2 min at 95 °C and a last cycle of 15 s at 60 °C. The cycle threshold (C_t) values were recorded and analysed using ViiA7 v1.2 software.

CopyRighter[51] v0.46 was applied to qPCR counts to correct for 16S copy number variation. CopyRighter normalizes the relative abundances across OTUs for each sample after dividing by the estimated copy number in a pre-computed table. The OTU genomic abundance is then obtained by multiplying by the total abundance number. A new CopyRighter database table was generated for the 2013 GreenGenes taxonomy (Supplementary Data 8), with copy number estimates for leaf OTUs as the average copy number of IMG version 4.1 genomes mapped to GreenGenes genomes and clustered at 99% sequence identity, and for higher taxonomic levels inferred copy numbers for the clade common ancestor. The inferred copy numbers for higher taxonomies were propagated to descendent lineages without known copy numbers.

SingleM. To determine microbiome diversity and community structure, SingleM was applied to reads from each sample (B.J.W. et al., unpublished materials, source code available at https://github.com/wwood/singlem).

Diversity calculations. Shannon diversity[52] was calculated based on SingleM counts, rarefying to 100 sequences per marker gene when >100 sequences were detected and excluding samples otherwise. Vegan[53] was used to calculate the diversity given the rarefied SingleM OTU table across each of the 15 marker genes, and the average was plotted in Fig. 1.

Genome assembly and binning. Each sample's reads were assembled individually using CLC Genomics Workbench version 4.4 (CLC Genomics) with an estimated insert size of 50–500, generating 214 assemblies. Differential coverage binning was undertaken by mapping all reads from each sample of site (palsa, bog or fen) to all assemblies of that site, using BamM 'make' (M. Imelfort, unpublished materials, http://ecogenomics.github.io/BamM/) version 1.3.8-1.5.0, BWA 0.7.12[54], samtools[55], and GNU parallel[56]. Each sample's scaffolds were then binned using MetaBAT 3127e20aa4e7[57] using the sample's contigs and each of the BAM files as points of differential coverage.

The CheckM[58] v1.0.4 'lineage_wf' pipeline was used to determine completeness and contamination of the MAG bins through the identification and quantification of single-copy marker genes, making use of pplacer 1.1 alpha 16[59]. Genomes estimated to be more than 70% complete and less than 10% contaminated were designated the 'Stordalen MAGs'.

MAG dereplication and taxonomic classification. When calculating relative abundance, to alleviate multi-mapping issues, genomes were first dereplicated at 97% average nucleotide identity (ANI). First, amino acid identity was calculated between all genomes using the CompareM (v0.0.17) AAI workflow ('comparem aai_wf', D. H. Parks, unpublished materials, https://github.com/dparks1134/CompareM). Genomes with an AAI of >95% were compared with each other using 'calculate_ani.py' (L. Pritchard, unpublished materials, https://github.com/widdowquinn/scripts). Genomes with >97% ANI over >70% alignment were clustered together using single-linkage clustering, and the genome with highest quality in each cluster was chosen as the representative, where quality was calculated as 'completeness – 4 × contamination', as estimated by CheckM above. The cluster representative for each recovered MAG is provided in Supplementary Data 3. The CompareM AAI workflow was also used to determine average amino acid identities between cluster representatives to determine the specific clades (Supplementary Note 7).

Genome tree and phylogenetic inference of 1,529 population bins. Phylogenetic inference was conducted in order to classify the MAG bins and used an in-house pipeline described in detail previously[60], the genome taxonomy database (GTDB v2.1.15)[92]. In brief, sets of 122 archaeal and 120 bacterial specific single-copy marker genes were used to infer domain-specific trees incorporating the 1,529 MAGs, a reference set of genomes from NCBI (RefSeq[61] release 80), and the recently published UBA genomes[62]. The concatenated alignment of these marker genes was created using HMMER v3.1.b2, and used as a basis for FastTree v2.1.9[63] tree building under the WAG + GAMMA model and using the approximately maximum likelihood method. This tree was then bootstrapped using genometreetk v0.0.35 (D. H. Parks, unpublished materials, https://github.com/dparks1134/GenomeTreeTk), calculating bootstrap support from 100 FastTree iterations. The associated taxonomy was derived using NCBI annotations, and was used to decorate the tree using tax2tree[64] and adjusted manually. Trees were visualized in ARB v6.0.6[65], and exported into ITOL[66] for further refinements before final editing in Inkscape. For the overall Bacteria and Archaea tree the dereplicated set of 647 genomes were selected in ARB and exported for viewing in ITOL. For the Acidobacteria tree (Extended Data Fig. 1b), Aminicenantes, including two Stordalen MAGs, and the recently reported *Ca.* 'Fischerbacteria'[9] were included as likely classes within the Acidobacteria based on GTDB analysis (http://gtdb.ecogenomic.org/). The bootstrapped Newick trees for the overall Bacteria and Archaea trees are found in Supplementary Data 10 and 11, using the alignments from Supplementary Data 12 and 13. Phylogenetic gain (Fig. 1, Extended Data Fig. 1) was calculated using genometreetk pd_clade, and based on the added phylogenetic distances introduced to current phyla (comprising RefSeq release 80 and UBA genomes) by including the 1,529 Stordalen MAGs.

Calculation of relative abundance. To calculate the relative abundance of each genome in each lineage, reads from each sample were mapped to the set of dereplicated genomes using BamM 'make'. Low quality mappings were removed with BamM v1.7.3 'filter' (minimum identity 95%, minimum aligned length 75% of each read) and the coverage of each contig calculated with BamM 'parse' using 'tpmean' mode, so calculating the coverage as the mean of the number of reads aligned to each position, after removing the highest 10% and lowest 10% of positions. The coverage of each MAG was calculated as the average of contig coverages, weighting each contig by its length in base pairs. The relative abundance of each lineage in each sample was calculated as its coverage divided by the total coverage of all genomes in the dereplicated set.

© 2018 Springer Nature Limited. All rights reserved.

Genomes that were differentially abundant by depth. To determine which lineages were differentially abundant between surface and deep samples, the set of relative abundances from each surface sample was compared to the set of relative abundances in the deep samples. The mean and statistical significance of the difference was calculated use R v3.3.2[67]. To determine the average amino acid identity between pairs of samples, the 'aai_wf' of CompareM v0.0.7 (D. H. Parks, unpublished materials, https://github.com/dparks1134/CompareM) was used using the protein sequences predicted by Prokka as input.

Annotation. Gene calling and preliminary annotation were undertaken with Prokka 1.11[68]. The genome was either annotated as Archaea or Bacteria, based in an inferred domain derived from the genome tree detailed above.

Annotation of glycoside hydrolase genes. All proteins predicted from all recovered genomes were screened using HMMSEARCH[69] using the dbCAN HMMs v5[70], using default parameters, then results post-processed to remove hits with $e > 1 \times 10^{-18}$ and HMM coverage of <0.35, where coverage was calculated as (hmm_to - hmm_from/qlen). Any genes with a hit passing these thresholds was then mapped to an EC number[71] using DIAMOND v0.8.27.89[72], with a database of all genes annotated with a fully defined (four number) E.C. number. This database of E.C. annotated genes was generated by gathering a list of GenBank identifiers of all characterized genes from each CAZy webpage[73] (listed at http://www.cazy.org/Glycoside-Hydrolases.html) using a custom Ruby script and then downloading the corresponding protein sequences from GenBank.

Annotation of carbon metabolism. Annotation was undertaken using in-house scripts, which assign KEGG orthology to each gene via HMMs, taking the best hit and requiring an $e < 1 \times 10^{-5}$. Encoding of whole pathways was inferred from genomes through the application of KEGG modules, both those available from KEGG as well as a number of custom modules (Supplementary Data 9).

Etymology. Description of '*Candidatus* Methanoflorens crillii' sp. nov.

'*Candidatus* Methanoflorens crillii' ['cril.lii. N. L. gen. n. '*crillii*', named after Patrick Crill, Stockholm University, Sweden, in recognition of his work on understanding of biogeochemical processes at the landscape scale (thawing permafrost) including greenhouse gases emission under the impact of climate change].

Candidatus Methanoflorens crillii sp. nov. is the second species recognized in the genus '*Candidatus* Methanoflorens'. The description is as provided by Mondav et al. (2014) for the genus with the following additional properties. The species can be differentiated from the recognized '*Ca*. M. stordalenmirensis' on the basis of phylogenetic analyses showing them to be monophyletic and sufficiently distinct average amino acid identity between encoded proteins.

Description of '*Candidatus* Acidiflorens stordalenmirensis' gen. et sp. nov.

'*Candidatus* Acidiflorens stordalenmirensis' (A.ci.di.flo´rens. N.L. n. acidum (from L. adj. acidus, sour), an acid; N.L. masc. substantive from L. masc. part. adj. florens, flourishing, to bloom; N. L. masc. n. Acidiflorens, an organism that blooms in acid; stor.da.len.mi.ren´sis. N.L. masc. adj. 'stordalenmirensis', of or belonging to Stordalen Mire, Sweden from where the species was characterized).

Description (brief). Phylogenetic analyses of genes/markers indicated that this species is different from all other recognized genera in the family Acidobacteriaceae.

Description of '*Candidatus* Changshengia' gen. nov.

'*Candidatus* Changshengia' (Chan.gshen´gia. N. L. fem. n. '*Candidatus* Changshengia', named in honour of Changsheng Li of The University of New Hampshire, a developer of the DeNitrification-DeComposition (DNDC) ecosystem model that contributed to our understating of the soil biogeochemical processes occurring in a variety of terrestrial ecosystems and climatic conditions).

Candidatus Changshengia gen. nov. is the second proposed and characterized genus in the phylum Dormibacteraeota. The delineation of genus is based on average amino acid identity between encoded proteins.

Description of '*Candidatus* Methanoflorentales' order nov.

'*Candidatus* Methanoflorentales' (N.L. masc. adj. '*Candidatus* Methanoflorens', type genus of the order; suff. –ales, ending to denote an order; N.L. fem. pl. n. '*Candidatus* Methanoflorentales' the order of the genus '*Candidatus* Methanoflorens').

The description is the same as given for the type genus '*Candidatus* Methanoflorens' and the family '*Candidatus* Methanoflorentaceae' Mondav et al. (2014) with the following modifications. The delineation of the order is determined by phylogenetic analyses showing that the Methanocellales would otherwise be paraphyletic. The order currently comprises two species '*Candidatus* M. stordalenmirensis' and '*Candidatus* M. crillii'. The type genus is '*Candidatus* Methanoflorens'.

Production and consumption rates of methane. Per-cell methane production and consumption rates were taken from studies of isolate cultures (for production[74–76] 0.19–4.5 fmol CH_4 $cell^{-1}$ h^{-1} and for consumption[77–79] 0.2–40 fmol CH_4 $cell^{-1}$ h^{-1}). Rates were taken as the mean of these rates for production and consumption rates, respectively.

Bulk density measurements. Data are from one palsa core sample taken from July 2013. In the field, 50 cm^3 aliquots of fresh peat from each core section were removed and frozen. In the laboratory, each 50 cm^3 section of peat was weighed, freeze dried and then reweighed. Bulk densities were determined gravimetrically and calculated from the freeze-dried weights of the volumetric sections. Water contents were determined by the per cent change in weight of the peat before and after freeze-drying.

Metatranscriptomics. Metatranscriptome sequencing was conducted on select samples from 2010, 2011 and 2012, comprising four palsa, eight bog and twelve fen samples. ScriptSeq Complete (Bacterial) low-input library preparation kits (Epicentre) were used with 240 ng of sample RNA that had been co-extracted alongside the DNA from the initial sample material as input as described previously. DNAse I (Roche) was used to remove residual DNA from the RNA after extractions. Agilent 2100 Bioanalyzer and Agilent 2200 Tapestation (Agilent Technologies) were used to check the quality of RNA and libraries during processing, with Qubit (ThermoFisher Scientific) used to determine quantity. These samples were run on 1/8th of a NextSeq (Illumina) lane, with initial shallow runs conducted on 1/11th of a HiSeq (Illumina) and MiSeq (Illumina) lanes. Files originating from the same metatranscriptome libraries were concatenated before analysis.

SeqPrep (J. A. St John, https://github.com/jstjohn/SeqPrep) was used to remove sequencing adaptors. PhiX contamination was removed by mapping the reads against the PhiX genome using BamM, and reads that aligned were removed. SortMeRNA v2.1[80] was used to remove non-coding RNA sequences (tRNA, tmRNA, 5S, 16S, 18S, 23S, 28S). To assign expression values to each gene, reads were first mapped in pairs to the dereplicated set of MAGs using BamM make, and filtered using BamM filter with cutoffs of 95% identity and 75% alignment. The count of reads mapped to each gene was calculated using DirSeq (https://github.com/wwood/dirseq, internally using BEDTools[81]) based on any overlap of forward reads with the open reading frame of the gene, tabulating the sense and antisense mappings independently. To avoid the potential for DNA contamination of the RNA libraries to provide a misleading interpretation of a gene being expressed, the number of reads mapping in the sense direction were compared to the number mapping in the antisense direction using a one-sided binomial test. Genes with a significantly more reads mapping in the sense direction ($P < 0.05$) were classified as 'expressed'. For each significantly expressed gene, the number of antisense reads was subtracted from the number of sense reads to correct for metagenome contamination. These normalized expression estimates were used to calculate the TPM score[82], using only protein coding genes (CDS regions defined in the Prokka annotated GFF files) in each sample.

Pathway expression was calculated as the average expression of the steps within a pathway. If a pathway step included an enzyme complex, the average expression of each subunit was used as the expression value of that step. If a reaction could be catalysed by more than one enzyme, or if multiple copies of an enzyme were encoded by a genome, then their summed expression was used as the expression value of that step.

Metaproteomics. *Protein extraction, purification, and digestion.* Metaproteome analysis was conducted on 22 samples from 2012, collected from the same cores and depths as material used for metagenomes and metatranscriptomes. Three metaproteomes were created by pooling replicate cores (3 × 3 replicates). Sample nomenclature denotes year and month, followed by habitat (*P* = palsa, S = bog, E = fen), core number (123 indicates replicate cores 1, 2 and 3 were pooled) and depth (surface = S, mid = M, deep = D, extra-deep = X). The 16 resulting metaproteomes were as follows: four palsa (20120600_P123M, 20120700_P3M, 20120700_P3D, 20120800_P2M) six bog (20120600_S123M, 20120600_S123D, 20120700_S2M, 20120700_S1D, 20120800_S1M, 20120800_S1X), and six fen samples (20120700_E3M, 20120700_E3D, 20120700_E2X, 20120800_E2M, 20120800_E2D, 20120800_E3D) (Supplementary Data 1). Proteins were extracted and digested using substantial modifications of methods developed previously for our site[34]. In brief, samples were thawed and 35 g of peat per sample was split equally into two 50 ml tubes, and sodium dodecyl sulphate (SDS)-resuspension buffer was added to a final volume of 30 ml. SDS-resuspension buffer (pH 8) was freshly prepared as: (1) an SDS buffer of 50 mM dithiothreitol (DTT) in 10 ml of 4% SDS, (2) a separate resuspension buffer of 50 mM trisaminomethane (Tris) Buffer (2.21 g Trizma-HCl and 4.36 g Trizma Base (Millipore Sigma)), 150 mM NaCl, 1 mM EDTA, and HPLC-grade water up to 1 l, (3) the 10 ml SDS buffer (warmed at 60 °C for 2 min) was mixed with 40 ml of resuspension buffer, and the final pH was adjusted to 8. Samples were vortexed for 10 min using a tabletop vortexer with adapters for 50 ml conical tubes, and then 10 g of 0.1 mM glass beads (Qiagen, Hilden, Germany) were added, followed by 30 min of vortexing. Samples were centrifuged at 3,000*g* for 20 min, the supernatant transferred to a new tube and centrifuged at 4,800*g* for 20 min. The supernatant was transferred to a new tube, to which 100% trichloroacetic acid (TCA) was added to a final concentration of 30%. Samples were shaken and then stored at 4 °C overnight.

Samples were centrifuged at 4,800*g* for 1 h 30 min at 4 °C, and then supernatant was decanted and pellets from the same sample were combined. The following steps were repeated three times: pellets were washed with 1 ml cold acetone, placed on ice for 5 min, vortexed briefly, centrifuged at 24,000*g* for 25 min at 4 °C, and then

© 2018 Springer Nature Limited. All rights reserved.

supernatant was removed. Pellets were dried under N_2 gas, and then 1–1.5 ml of denaturing buffer was added. Denaturing buffer was prepared as follows: (1) a digestion buffer was prepared with 4.88 g of Trizma-HCl, 2.30 g of Trizma Base, and 1.11 g of $CaCl_2$ brought to a 1 l volume with HPLC-grade water, (2) guanidine-HCl was added to digestion buffer to a final concentration of 6 M in a 50-ml tube. Samples were incubated at 60 °C for 1 h, vortexing for 5 s every 10 min, then transferred to a new tube, to which digestion buffer was added to a final volume of 15 ml. Proteins were digested by adding 20 μg trypsin (NEB) and incubating on a nutating mixer at 37 °C overnight.

A further 10 μg of trypsin was added to each sample, followed by incubation on a nutating mixer at 37 °C for 4 h. DTT was added (approximately 100 mg), then samples were returned to the 37 °C nutating mixer for 45 min and then centrifuged at 3,000 r.p.m. for 5 min at 20 °C. The supernatant was transferred to a new tube. Similar to previously reported peptide purification protocols[83], Sep-Pak Plus C18 cartridges (Waters) were conditioned with 10 ml of acetonitrile + 0.1% formic acid, then washed with 10 ml of 0.1% formic acid (in water), then the samples were added and the flow-through was refiltered through the cartridges three additional times, followed by a wash with 10 ml of 0.1% formic acid (in water). Peptides were eluted with 5 ml of acetonitrile + 0.1% formic acid, then added to 0.45 μm Ultrafree-MC filter tubes according to the manufacturer's protocol (Millipore Sigma).

Peptide fractionation and mass spectroscopy. A bicinchoninic acid assay (Thermo Scientific, Rockford, IL) was performed to determine the peptide mass in each sample. The samples were then diluted with 10 mM ammonium formate, pH 10 ('buffer A') to a volume of 930 μl, centrifuged at 10,000 *g* for 2–5 min to remove any precipitates, and transferred to snap-cap ALS vials. The diluted samples (pH 10) were resolved on a XBridge C18, 250 × 4.6 mm, 5 μM with 4.6 × 20 mm guard column (Waters). Separations were performed at 0.5 ml/min using an Agilent 1100 series HPLC system (Agilent Technologies, Santa Clara, CA) with mobile phases (A) buffer A and (B) buffer A/acetonitrile (10:90). The gradient was adjusted from 100% A to 95% A over the first 10 min, 95% A to 65% A over minutes 10 to 70, 65% A to 30% A over minutes 70 to 85, maintained at 30% A over minutes 85 to 95, re-equilibrated with 100% A over minutes 95 to 105, and held at 100% A until minute 120. Fractions were collected every 1.25 min (96 fractions over the entire gradient) and every 12th fraction were pooled for a total of 12 fractions per sample. All fractions were dried under vacuum and 20 μl of nanopure water was added to each fraction for storage at −20 °C until LC–MS/MS analysis.

Fractions were analysed by reversed-phase LC–MS/MS using a Waters nanoAquity UPLC system coupled with a Q-Exactive Plus hybrid quadrupole/Orbitrap mass spectrometer from Thermo Fisher Scientific. The analytical column was packed in-house by slurry packing 3-μm Jupiter C_{18} stationary phase (Phenomenex) into a 70-cm long, 360 μm OD × 75 μm ID fused silica capillary tubing (Polymicro Technologies Inc.). Mobile phases consisted of 0.1% formic acid in water (MP-A) and 0.1% formic acid in acetonitrile (MP-B). Samples were adjusted to a concentration of ~ 0.1 μg/μl and 5 μl injections were directly loaded onto the analytical column at a flow rate of 300 nl/min and 1% MP-B. The full loading, gradient elution, and column regeneration profile was as follows (min:%MP-B); 0:1, 30:1, 32:8, 50:12, 105:30, 110:45, 120:90, 125:90, 130:1, 170:1. Data acquisition (100 min) was started at the end of the sample loading period (30 min). The analytical column was coupled to the Q-Exactive using a home-built nanospray adaptor interface with 2.2 kV applied to achieve electrospray ionization. The MS inlet was maintained at a temperature of 300 °C. A precursor scan was performed from m/z 300 to 1800 at a resolution of 30k and an automatic gain control (AGC) of 3×10^6. Operated in data dependent mode, the top 12 most intense ions from the precursor scan were selected for high energy collision dissociation (HCD) MS/MS at a resolution of 17.5k, AGC of 1e5, isolation window of 2 m/z, and a max ion time of 100 ms. Only ions identified as having a +2 charge or higher were subjected to HCD and subsequently excluded from further analysis for 30 s thereby allowing for deeper coverage. In total, 192 mass spectra files were generated (12 fractions for each of the 16 samples).

Database search and expression analysis. A sensitive and universal database search tool, MSGFPlus[84] v2017.01.13, was used to conduct the metaproteome searches in this study. Prior to searching the metaproteomes, the mass spectrometer RAW output files were converted to the mzML format using msConvert of ProteoWizard[85] 3.0.10200, accepting the default parameters. The mzML files were then searched against a targeted protein database containing protein sequences predicted from the metagenome-assembled genomes across the permafrost thaw gradient and involved in metabolic pathways examined in the manuscript, as well as the entire CDS regions of *Ca.* 'Acidiflorens', *Ca.* 'Methanoflorens' and AD3 (Supplementary Data 14). Proteins in the targeted database were dereplicated at 100% amino acid identity using usearch v9.2.64 (–fastx_uniques)[86] after converting all isoleucine residues to leucine (due to identical masses). In order to calculate the false discovery rate (FDR), a parallel search of a decoy protein database was conducted by using the (-tda 1) parameter of MSGFPlus during the indexing and searching steps of the targeted database. After conducting the searches, the FDR was calculated as: FDR(t) = #DecoyPSMs at (t)/#TargetPSMs at (t), where (t) is the highest Q value that gives a FDR of ≤ 1%, and #DecoyPSMs and #TargetPSMs are the numbers of Decoy and Target Peptide-Spectrum Matches (PSM), respectively, at that Q-value threshold. Only PSMs with $Q \leq (t = 0.0145)$ were considered in the results of this study. The maximum precursor mass tolerance allowed during the searches was specified by the (-t 20 ppm) parameter for parent mass tolerance and the (-ti '-1,2') parameter for isotope error range. Trypsin was specified as the digestion method (-e 1) and only full tryptic digestion was allowed (-ntt 2). Minimum and maximum peptide lengths to consider were 6 and 50, respectively. Minimum and maximum precursor charges to consider were 2 and 5, respectively. For each spectral scan, only the PSMs with the highest MSGF score (-n 1) were considered for subsequent analyses.

The detection of a unique peptide in at least one sample was considered evidence of expression of a specific enzyme. A specific pathway was designated as expressed when at least one enzyme from a pathway was detected in the metaproteome. Further, that enzyme had to be encoded in a genome where >50% of genomes in the 97% ANI genome cluster encoded all steps of that pathway. Only a very small number of proteins (<1%) were detected where the sequence of that protein was 100% identical to a protein from a different ANI genome cluster.

The limited number of proteins detected for the fen hydrolysers is likely a consequence of the diversity and complexity of the fen community, and the variety of proteins produced, compared to the palsa and bog. The metaproteomes were searched against a targeted protein database containing only those sequences predicted from the MAGs. Peptide-spectrum matching requires 100% matches, which means that variations in sequence due to population heterogeneity/diversity, which is greatest in the fen (Fig. 1; Extended Data Fig. 2), are not captured, reducing the number of successful matches. Our approach is conservative following the genome-centric focus, and used purely to show that the populations of interest are translating proteins in situ.

Statistical analyses. Analysis of variance (ANOVA) tests, Mann–Whitney *U*-tests, and least squares regressions were calculated using R[67]. Box plots were created using ggplot2[87], with the boxes representing the 1st to 3rd quartiles and the whiskers the highest (upper whiskers) or lowest (lower whiskers) observation within 1.5× the interquartile range. The centre line represents the mean. Figure 2 sample numbers were as follows: palsa surface $n = 18$, palsa mid $n = 18$, palsa deep $n = 17$, bog surface $n = 19$, bog mid $n = 20$, bog deep $n = 20$, bog extra-deep $n = 6$, fen surface $n = 23$, fen mid $n = 18$, fen deep $n = 23$, fen extra-deep $n = 6$ biologically independent samples. A square root scale was used on the y axis. P values for significant differences among depths assessed with ANOVA were 4×10^{-7}, 2×10^{-5}, and 2×10^{-5} for palsa, bog and fen respectively for cellulose degradation, and 4×10^{-2}, 3×10^{-8} and not applicable (n/a, no significant differences were found) for xylan degradation, 5×10^{-3}, n/a, n/a for xylose degradation (dehydratase), n/a, 2×10^{-12}, 1×10^{-4} for xylose degradation (oxidoreductase), 3×10^{-4}, 9×10^{-7}, 6×10^{-3} for xylose degradation (isomerase), 2×10^{-4}, 5×10^{-9}, n/a for lactate fermentation, n/a, 3×10^{-2}, n/a for ethanol fermentation, 8×10^{-5}, 5×10^{-7}, n/a for propionate fermentation, n/a, 2×10^{-4}, 5×10^{-2} for hydrogenotrophic methanogenesis, n/a, 2×10^{-3}, 4×10^{-3} for acetoclastic methanogenesis. Figure 3 sample numbers were as follows: 3a, $n = 47$, 3b, $n = 65$, 3c, $n = 70$ biologically independent samples.

To test for associations between microbial populations and geochemical variables, the abundance of genus- to family- level lineages was calculated using phylogenetic tree insertion of open reading frames derived from individual reads into single-copy ribosomal protein marker trees using GraftM[88] and phylogenetic trees derived from both contigs assembled from data presented here and reference datasets (data not shown). To avoid statistical complications arising from multiple hypothesis testing, only the ten most abundant lineages in the bog and fen were tested for significance using least squares regression. Lineages correlating significantly were then linked to MAGs for correlations between individual MAGs and geochemical variables as well as between *Ca.* 'Acidiflorens' sub-lineages.

Figure generation. Manuscript figures were generated using custom R[67] scripts, ggplot2[87], spacemacs (http://spacemacs.org/), Rstudio[89], arb[65], d3js (https://d3js.org/), Inkscape (https://inkscape.org/) and Adobe Illustrator (http://www.adobe.com/au/products/illustrator.html).

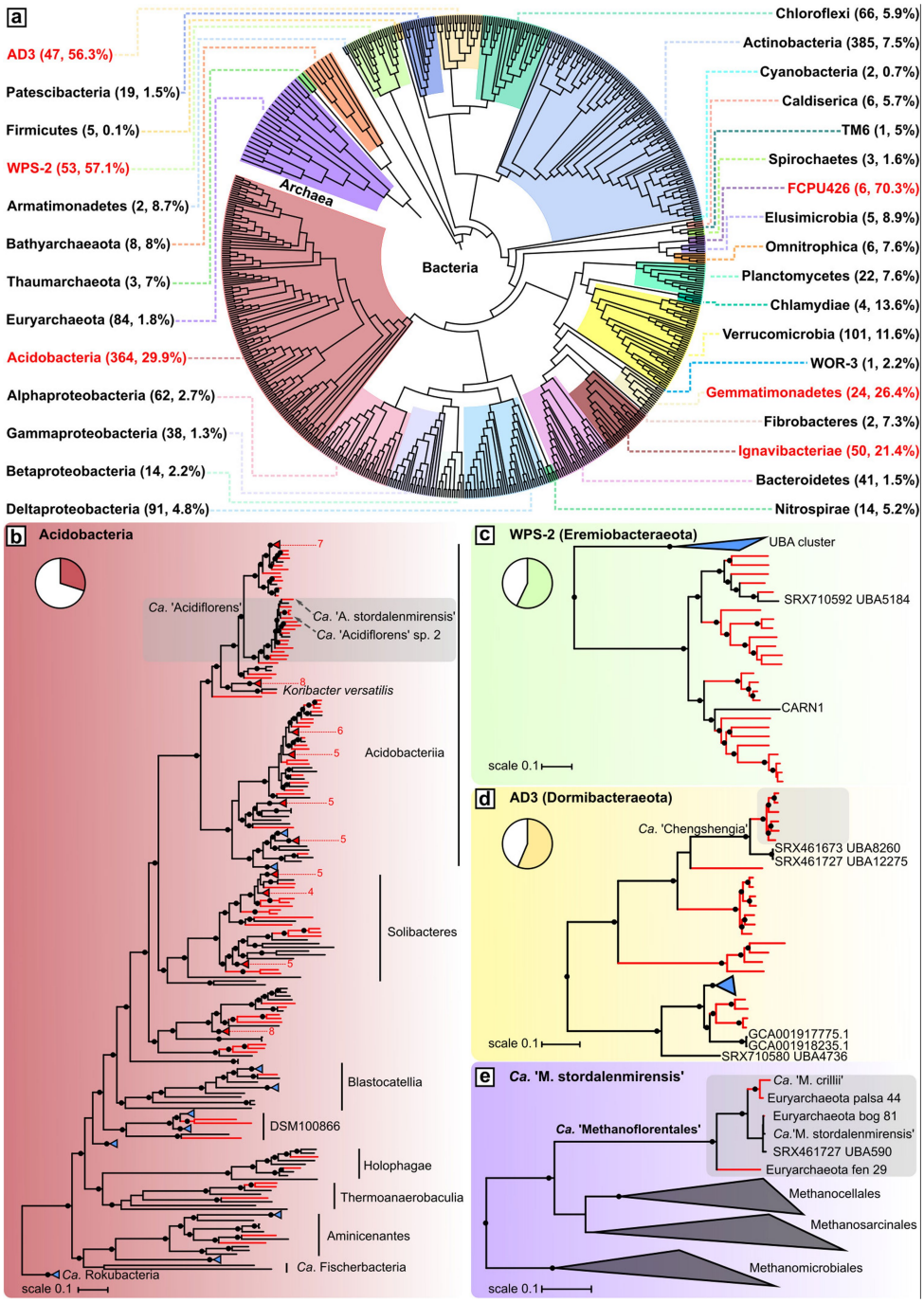
Code availability. The above methods indicate the source of the code and programs used for analyses within the relevant sections.

Data availability. Data described in this manuscript have been submitted under NCBI BioProject accession number PRJNA386568. MAGs were deposited at DDBJ/ENA/GenBank under the accession numbers provided in Supplementary Data 3, and the initial versions are described in this paper. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://www.proteomexchange.org/) via the PRIDE[90] partner repository with the dataset identifier PXD009096 (https://doi.org/10.6019/PXD009096). Supplementary Data 1–9 are available with the online version of this manuscript. Supplementary Data 10–15 are available on figshare (https://doi.org/10.6084/m9.figshare.6233660).

© 2018 Springer Nature Limited. All rights reserved.

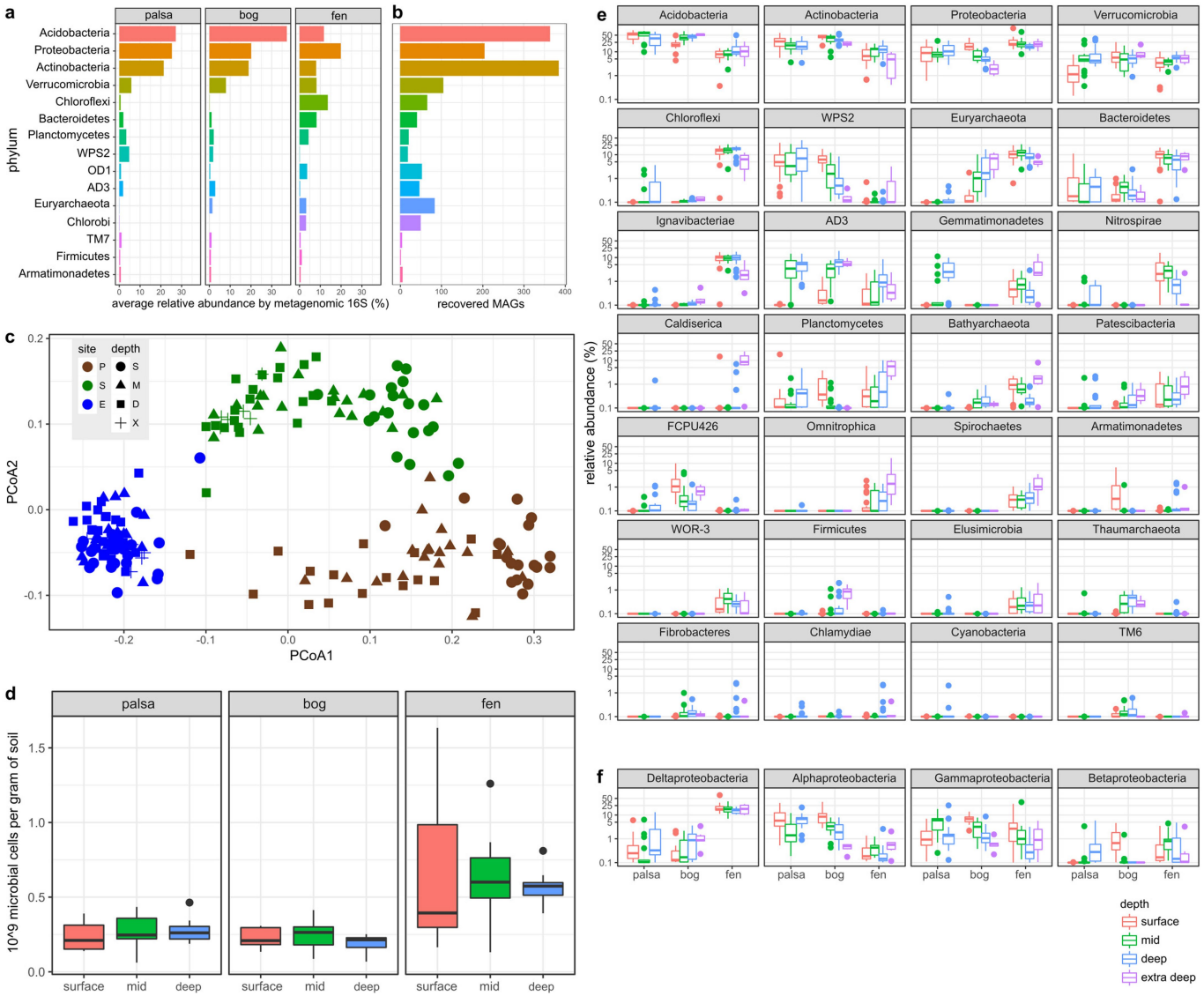
42. Vitt, D. H., Halsey, L. A. & Zoltai, S. C. The changing landscape of Canada's western boreal forest: the current dynamics of permafrost. *Can. J. For. Res.* **30**, 283–287 (2000).
43. Jorgenson, M. T., Racine, C. H., Walters, J. C. & Osterkamp, T. E. Permafrost degradation and ecological changes associated with a warming climate in central Alaska. *Clim. Change* **48**, 551–579 (2001).
44. Payette, S., Delwaide, A., Caccianiga, M. & Beauchemin, M. Accelerated thawing of subarctic peatland permafrost over the last 50 years. *Geophys. Res. Lett.* **31**, L18208 (2004).
45. O'Donnell, J. A. et al. The effects of permafrost thaw on soil hydrologic, thermal, and carbon dynamics in an Alaskan peatland. *Ecosystems (N. Y.)* **15**, 213–229 (2012).
46. Zoltai, S. Cyclic development of permafrost in the peatlands of northwestern Alberta, Canada. *Arct. Alp. Res.* **25**, 240–246 (1993).
47. Quinton, W., Hayashi, M. & Chasmer, L. Permafrost-thaw-induced land-cover change in the Canadian subarctic: implications for water resources. *Hydrol. Processes* **25**, 152–158 (2011).
48. Koven, C. D. et al. Permafrost carbon–climate feedbacks accelerate global warming. *Proc. Natl Acad. Sci. USA* **108**, 14769–14774 (2011).
49. Melton, J. et al. Present state of global wetland extent and wetland methane modelling: conclusions from a model intercomparison project (WETCHIMP). *Biogeosciences* **10**, 753–788 (2013).
50. Whiticar, M. J., Faber, E. & Schoell, M. Biogenic methane formation in marine and freshwater environments: CO_2 reduction vs. acetate fermentation—isotope evidence. *Geochim. Cosmochim. Acta* **50**, 693–709 (1986).
51. Angly, F. E. et al. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* **2**, 11 (2014).
52. Pielou, E. C. Shannon's formula as a measure of specific diversity: its use and misuse. *Am. Nat.* **100**, 463–465 (1966).
53. Oksanen, J. et al. The vegan package. *Commun. Ecol. Package* **10**, 631–637 (2007).
54. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
55. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
56. Tange, O. GNU parallel—the command-line power tool. *The USENIX Magazine* **36**, 42–47 (2011).
57. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
58. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
59. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).
60. Evans, P. N. et al. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).
61. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
62. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
63. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
64. McDonald, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).
65. Ludwig, W. et al. ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
66. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
67. R Computing Team *R Language Definition* (R Foundation for Statistical Computing, Vienna, 2000).
68. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
69. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
70. Yin, Y. et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–W51 (2012).
71. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).
72. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
73. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
74. Morozova, D. & Wagner, D. Stress response of methanogenic archaea from Siberian permafrost compared with methanogens from nonpermafrost habitats. *FEMS Microbiol. Ecol.* **61**, 16–25 (2007).
75. Lay, J.-J., Miyahara, T. & Noike, T. Methane release rate and methanogenic bacterial populations in lake sediments. *Water Res.* **30**, 901–908 (1996).
76. Li, Y.-Y. & Noike, T. Upgrading of anaerobic digestion of waste activated sludge by thermal pretreatment. *Water Sci. Technol.* **26**, 857–866 (1992).
77. Hanson, R. S. & Hanson, T. E. Methanotrophic bacteria. *Microbiol. Rev.* **60**, 439–471 (1996).
78. Costello, A. M., Auman, A. J., Macalady, J. L., Scow, K. M. & Lidstrom, M. E. Estimation of methanotroph abundance in a freshwater lake sediment. *Environ. Microbiol.* **4**, 443–450 (2002).
79. Baani, M. & Liesack, W. Two isozymes of particulate methane monooxygenase with different methane oxidation kinetics are found in *Methylocystis* sp. strain SC2. *Proc. Natl Acad. Sci. USA* **105**, 10203–10208 (2008).
80. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
81. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
82. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
83. Guo, X. & Kristal, B. S. The use of underloaded C(18) solid-phase extraction plates increases reproducibility of analysis of tryptic peptides from unfractionated human plasma. *Anal. Biochem.* **426**, 86–90 (2012).
84. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
85. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
86. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
87. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, Heidelberg, 2016).
88. Boyd, J. A., Woodcroft, B. J. & Tyson, G. W. GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. *Nucleic Acids Res.* **46**, e59 (2018).
89. Racine, J. S. RStudio: A platform-independent IDE for R and Sweave. *J. Appl. Econ.* **27**, 167–172 (2012).
90. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44** (D1), D447–D456 (2016).
91. Bertin, P. N. et al. Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics. *ISME J.* **5**, 1735–1747 (2011).
92. Parks, D. H. et al. A proposal for a standardized bacterial taxonomy based on genome phylogeny. Preprint at https://www.biorxiv.org/content/early/2018/01/30/256800 (2018).

© 2018 Springer Nature Limited. All rights reserved.



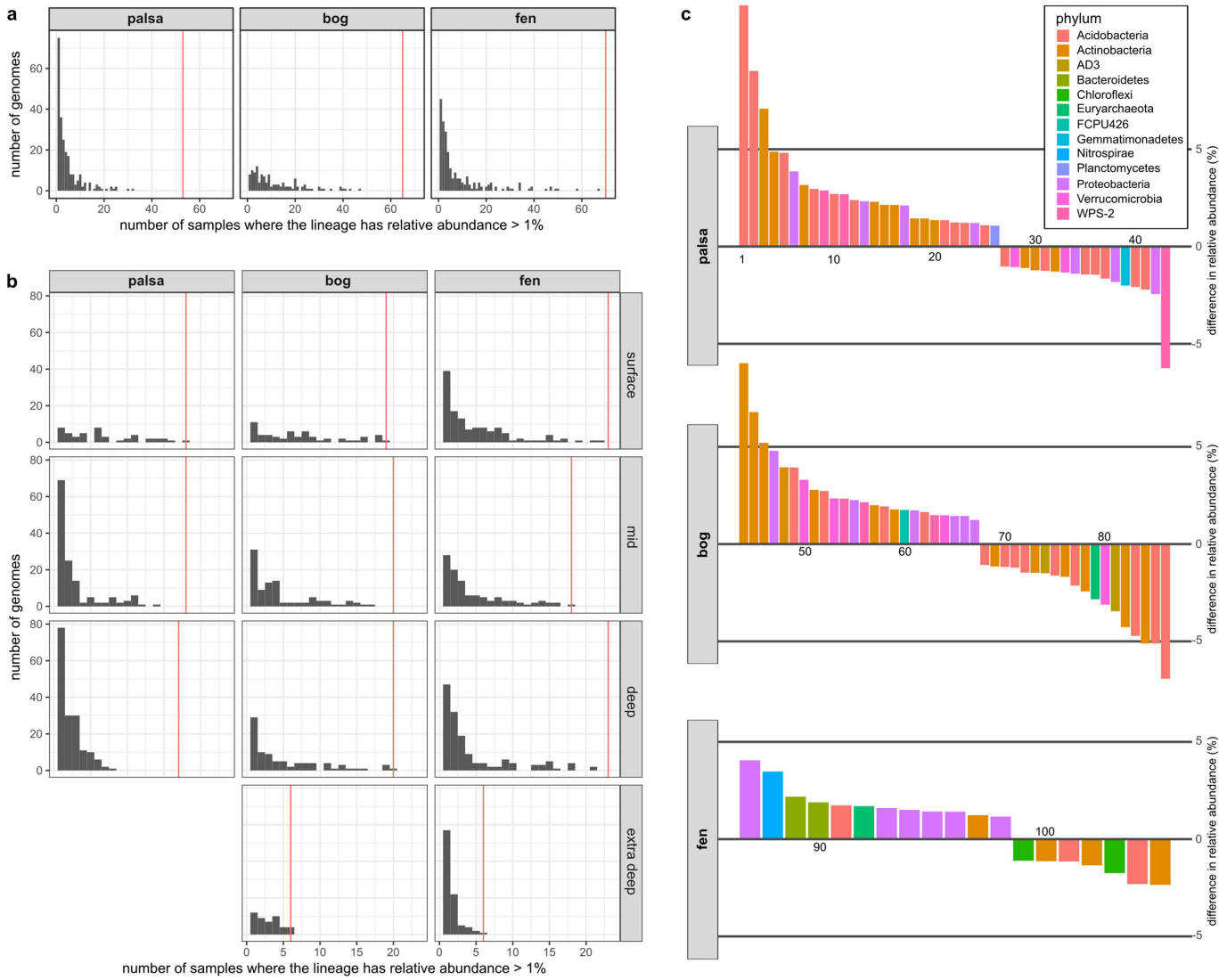
Extended Data Fig. 1 | Phylogenetic distribution of MAGs recovered from Stordalen Mire. a, Phylogenetic tree of 647 dereplicated MAGs. Numbers in parentheses show total MAGs recovered and phylogenetic gain of Stordalen MAGs compared to publicly available genomes for each phylum. Red text indicates previously poorly represented phyla. **b**, Acidobacteria subtree showing the *Ca.* 'Acidiflorens' lineages. **c**, Eremiobacteraeota subtree incorporating the CARN1[91] MAG. **d**, Dormibacteraeota subtree, showing *Ca.* 'Changshengia'. **e**, Subtree of *Ca.* 'Methanoflorentales' MAGs, and closest neighbouring orders. In **b**–**e**, pie charts show phylogenetic gain, red lines indicate Stordalen MAGs, black lines indicate public genomes, blue triangles indicate clustered public genomes and red triangles indicate clustered Stordalen MAGs. Black dots indicate bootstrap values 70–100%.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 2 | Microbial community profile of the thaw gradient. a, Relative abundance of each phylum estimated through the recovery of 16S rRNA gene reads, averaged within each thaw stage. The 15 phyla with the highest relative abundance across all samples are shown. **b**, Number of MAGs recovered from each of the phyla in **a**, showing that broadly, MAGs recovered are from lineages highest in abundance. **c**, Principal coordinates analysis of weighted UniFrac compositional differences between samples, based on average coverage of each recovered genome of reads mapped to the dereplicated genome set. Colours indicate thaw stage: brown = palsa (P), green = sphagnum/bog (S), blue = eriophorum/fen (E). Depth: S = surface, M = mid-depth, D = deep, X = extra-deep. Goodness of fit was 0.57 for PCoA 1 and 0.65 for PCoA 2. Sample numbers: $n = 53$, 65 and 70 biologically independent samples for palsa, bog and fen, respectively. **d**, Quantitative PCR analysis of samples taken in 2012. The number of cells per gram of soil is shown for three depths at the three thaw stages, after correcting for 16S rRNA gene copy number variation (see Methods). Fen samples contained significantly more cells per gram of soil than bog and palsa samples (average 2.6×, $P = 7 \times 10^{-8}$, $n = 103$, two-sided Mann–Whitney U-test). Sample numbers: $n = 8$, 9, and 8 for biologically independent samples palsa surface, mid and deep, respectively, $n = 9$, 8, 9 and 10, 7 and 9 for bog and fen, respectively. **e**, **f**, Relative abundances of phyla and classes within the Proteobacteria across the thaw gradient, respectively. The depth of each sample is indicated by the colour of the box (surface: red, mid-depth: green, deep: blue, extra-deep: purple). Each data point is the sum of relative abundances of all lineages assigned to the phylum in a sample after adding a 0.1% pseudocount to all phyla (so the y axis is not dominated by small values visually). Box plots are shown plotted on a log-scale y axis, with phyla and classes ordered by decreasing average relative abundance across all samples. Relative abundance was calculated based on the fraction of the community with recovered genomes (see Methods). Sample numbers: $n = 53$, 65 and 70 biologically independent samples for palsa, bog and fen, respectively.

© 2018 Springer Nature Limited. All rights reserved.

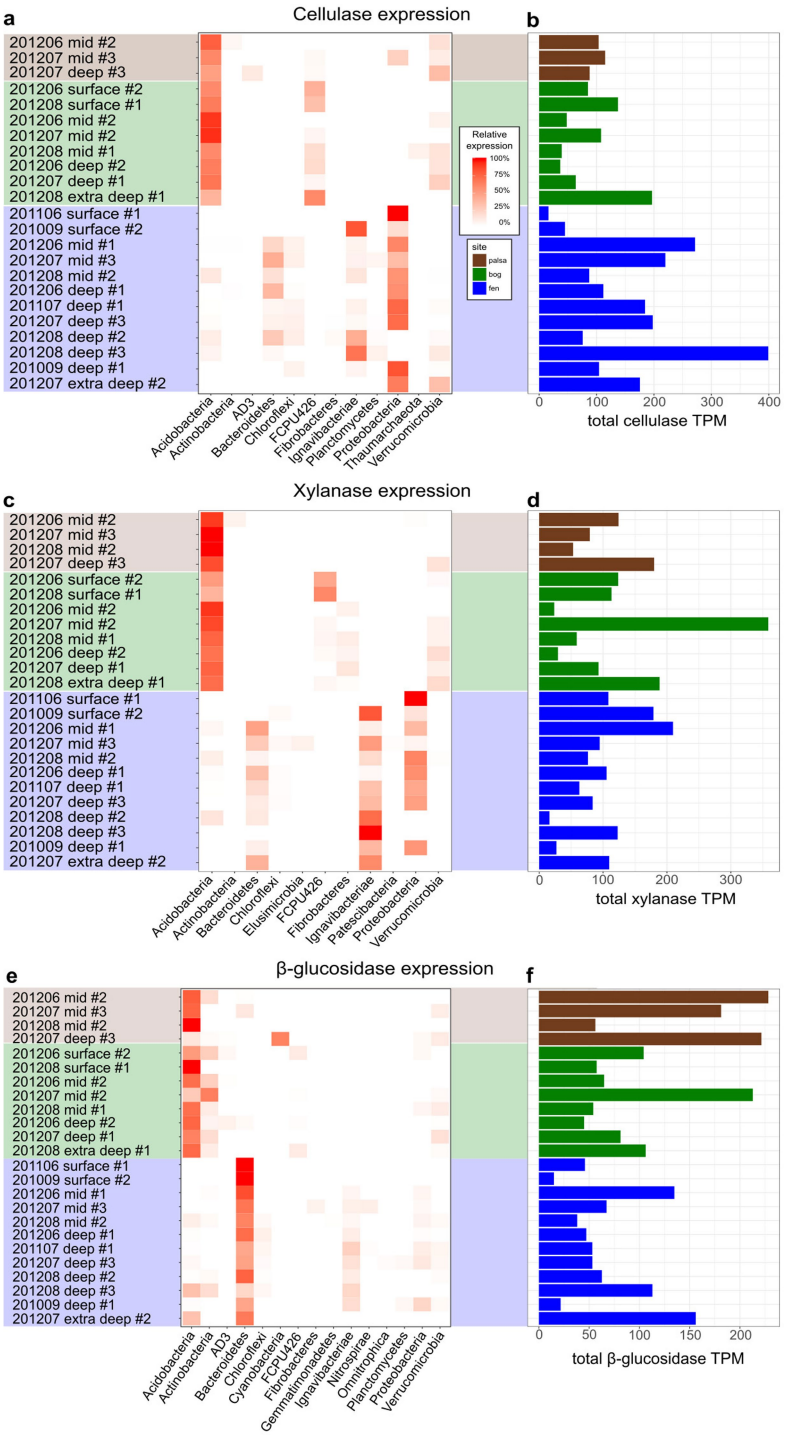


Extended Data Fig. 3 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

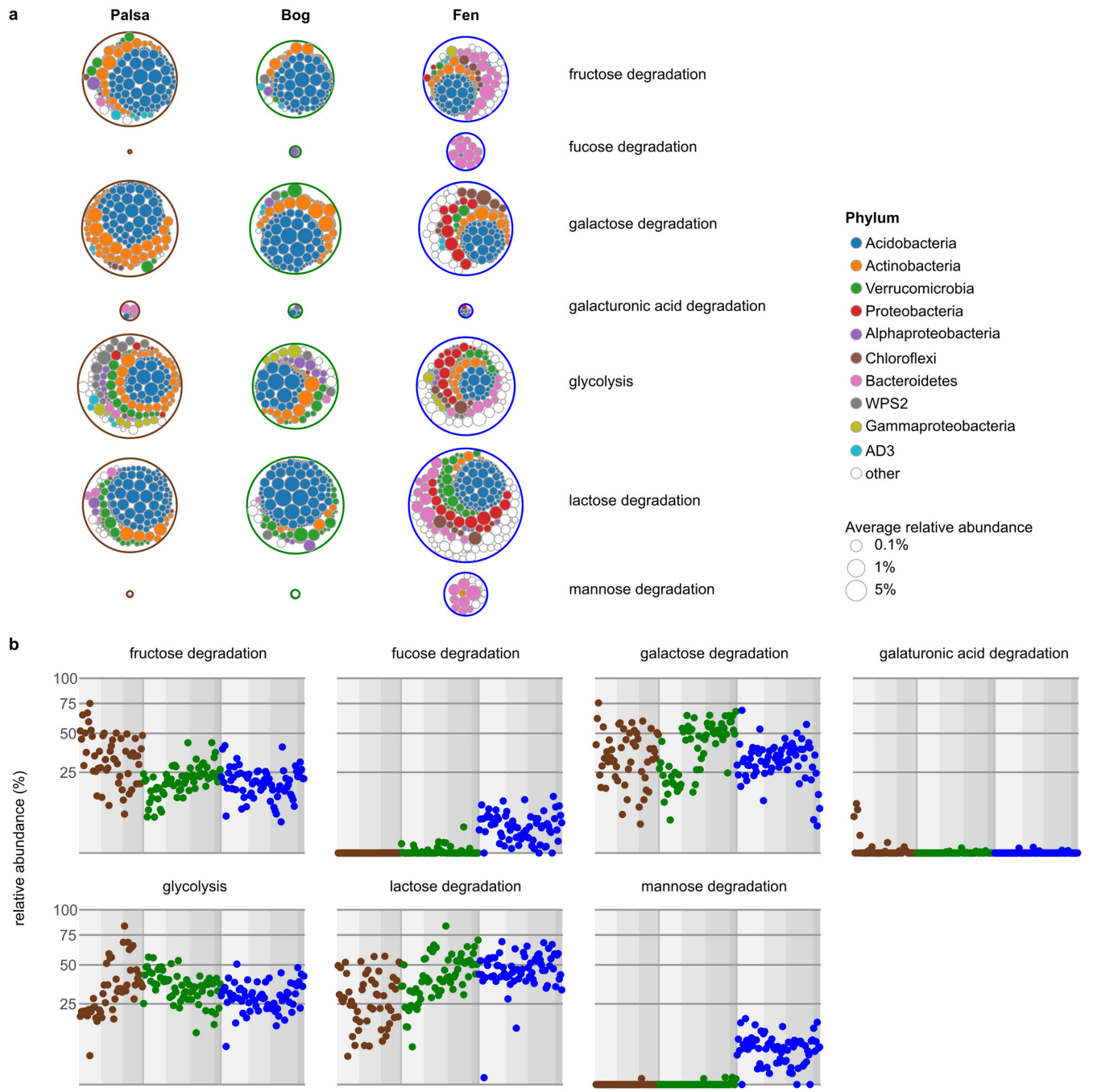
Extended Data Fig. 3 | Prevalence of individual MAGs across the thaw gradient. a, Number of samples where each Stordalen MAG is present at >1% relative abundance among each stage of the thaw gradient. Vertical red lines indicate the number of samples sequenced in total from that environment. Only one MAG, 'Deltaproteobacteria_fen_1087', was found in a high abundance across fen sites, detected at >1% relative abundance in 96% of fen sites. **b**, The same information stratified by depth of the sample in the soil column. The specific MAGs prevalent are detailed in Extended Data Table 1, showing that a small number of populations were prevalent at a specific depth of a specific site. **c**, Stordalen genomes that changed significantly in abundance with depth. For each site, genomes that show the largest absolute difference in abundance between shallow and deep samples are shown. Genomes that are more abundant in shallow samples compared to deep are positive, and those more abundant in deep samples relative to shallow samples are negative. Only those lineages with a mean absolute difference of >1% and that are significantly different ($P < 0.05$, two-sided Mann–Whitney U-test) are shown. Sample numbers: $n = 53$, 65 and 70 biologically independent samples for palsa, bog and fen, respectively. Each bar indicates a 97% dereplicated MAG that changes in relative abundance between surface and deep samples and the colour of each indicates the phylum the genome belongs to. The fen is less stratified between the surface and deep, which is reflected in the fewer population abundances significantly changing in abundance between shallow and deep samples. Recovered congeneric genomes that showed significant but inverse differential abundance between surface and deep samples are shown in Supplementary Data 7. Genomes depicted in **c** in order are Acidobacteria_palsa_348 = 1, Acidobacteria_palsa_246 = 2, Actinobacteria_palsa_463 = 3, Actinobacteria_palsa_558 = 4, Acidobacteria_palsa_312 = 5, Alphaproteobacteria_palsa_929 = 6, Actinobacteria_palsa_504 = 7, Acidobacteria_palsa_125 = 8, WPS2_palsa_1515 = 9, Acidobacteria_palsa_289 = 10, WPS2_palsa_1516 = 11, Acidobacteria_palsa_310 = 12, Alphaproteobacteria_palsa_913 = 13, Actinobacteria_palsa_693 = 14, Actinobacteria_palsa_465 = 15, Actinobacteria_palsa_691 = 16, Alphaproteobacteria_palsa_895 = 17, Actinobacteria_palsa_505 = 18, Actinobacteria_bog_593 = 19, Actinobacteria_palsa_462 = 20, Acidobacteria_palsa_199 = 21, Acidobacteria_palsa_362 = 22, Acidobacteria_palsa_313 = 23, Gammaproteobacteria_palsa_1209 = 24, Acidobacteria_palsa_267 = 25, Planctomycetes_palsa_1347 = 26, Acidobacteria_palsa_143 = 27, Verrucomicrobia_palsa_1397 = 28, Actinobacteria_palsa_641 = 29, Actinobacteria_palsa_733 = 30, Acidobacteria_palsa_420 = 31, Actinobacteria_palsa_736 = 32, Verrucomicrobia_palsa_1413 = 33, Alphaproteobacteria_palsa_910 = 34, Acidobacteria_palsa_286 = 35, Acidobacteria_palsa_122 = 36, Acidobacteria_palsa_343 = 37, Deltaproteobacteria_palsa_1114 = 38, Gemmatimonadetes_palsa_1248 = 39, Acidobacteria_palsa_340 = 40, Acidobacteria_palsa_141 = 41, Alphaproteobacteria_palsa_922 = 42, WPS2_palsa_1496 = 43, Actinobacteria_bog_635 = 44, Actinobacteria_bog_766 = 45, Actinobacteria_bog_592 = 46, Gammaproteobacteria_bog_1200 = 47, Actinobacteria_bog_594 = 48, Acidobacteria_bog_329 = 49, Verrucomicrobia_bog_1475 = 50, Actinobacteria_bog_723 = 51, Acidobacteria_bog_233 = 52, Verrucomicrobia_bog_1402 = 53, WPS2_bog_1492 = 54, Alphaproteobacteria_bog_899 = 55, WPS2_bog_1527 = 56, Actinobacteria_bog_769 = 57, Acidobacteria_bog_377 = 58, Actinobacteria_bog_637 = 59, FCPU426_bog_1183 = 60, Alphaproteobacteria_bog_900 = 61, Acidobacteria_bog_234 = 62, WPS2_bog_1502 = 63, Verrucomicrobia_bog_1421 = 64, Gammaproteobacteria_bog_1206 = 65, Alphaproteobacteria_bog_908 = 66, Betaproteobacteria_bog_994 = 67, Acidobacteria_fen_416 = 68, Actinobacteria_fen_548 = 69, Acidobacteria_bog_445 = 70, Acidobacteria_bog_96 = 71, Acidobacteria_bog_202 = 72, Actinobacteria_fen_455 = 73, AD3_bog_854 = 74, Acidobacteria_bog_218 = 75, Actinobacteria_bog_806 = 76, Acidobacteria_bog_390 = 77, Actinobacteria_bog_524 = 78, Euryarchaeota_bog_81 = 79, Verrucomicrobia_bog_1459 = 80, AD3_bog_876 = 81, Actinobacteria_bog_808 = 82, Acidobacteria_bog_226 = 83, Actinobacteria_bog_576 = 84, Acidobacteria_bog_406 = 85, Acidobacteria_fen_408 = 86, Deltaproteobacteria_fen_1088 = 87, Nitrospirae_fen_1304 = 88, Bacteroidetes_fen_982 = 89, Bacteroidetes_fen_956 = 90, Acidobacteria_fen_335 = 91, Euryarchaeota_fen_63 = 92, Gammaproteobacteria_fen_1191 = 93, Deltaproteobacteria_fen_1087 = 94, Gammaproteobacteria_fen_1218 = 95, Gammaproteobacteria_fen_1219 = 96, Actinobacteria_fen_730 = 97, Deltaproteobacteria_fen_1138 = 98, Chloroflexi_fen_1050 = 99, Actinobacteria_fen_453 = 100, Acidobacteria_fen_408 = 101, Actinobacteria_fen_548 = 102, Chloroflexi_fen_1019 = 103, Acidobacteria_fen_414 = 104, Actinobacteria_fen_455 = 105.

© 2018 Springer Nature Limited. All rights reserved.



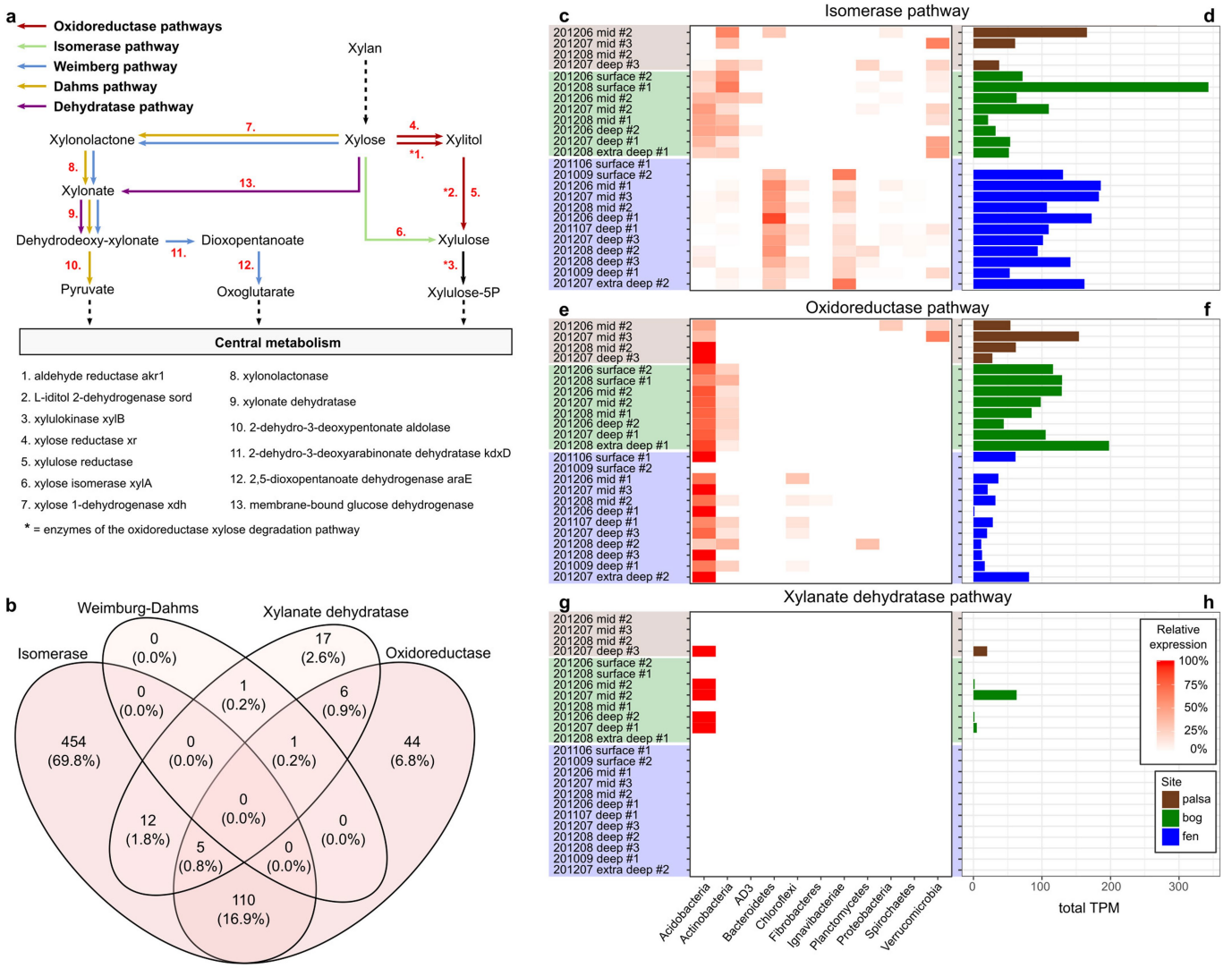
Extended Data Fig. 4 | Cellulase, xylanase and β-glucosidase gene expression across the thaw gradient. a, **b**, Cellulose; **c**, **d**, xylanase; **e**, **f**, β-glucosidase. Samples analysed with metatranscriptomics are described by the date of sampling, core number and depth. **a**, **c**, **e**, Relative contribution of each phylum to the total TPM of the enzyme class observed in the metatranscriptomes. **b**, **d**, **f**, Total TPM of all expressed genes in the sample.

© 2018 Springer Nature Limited. All rights reserved.



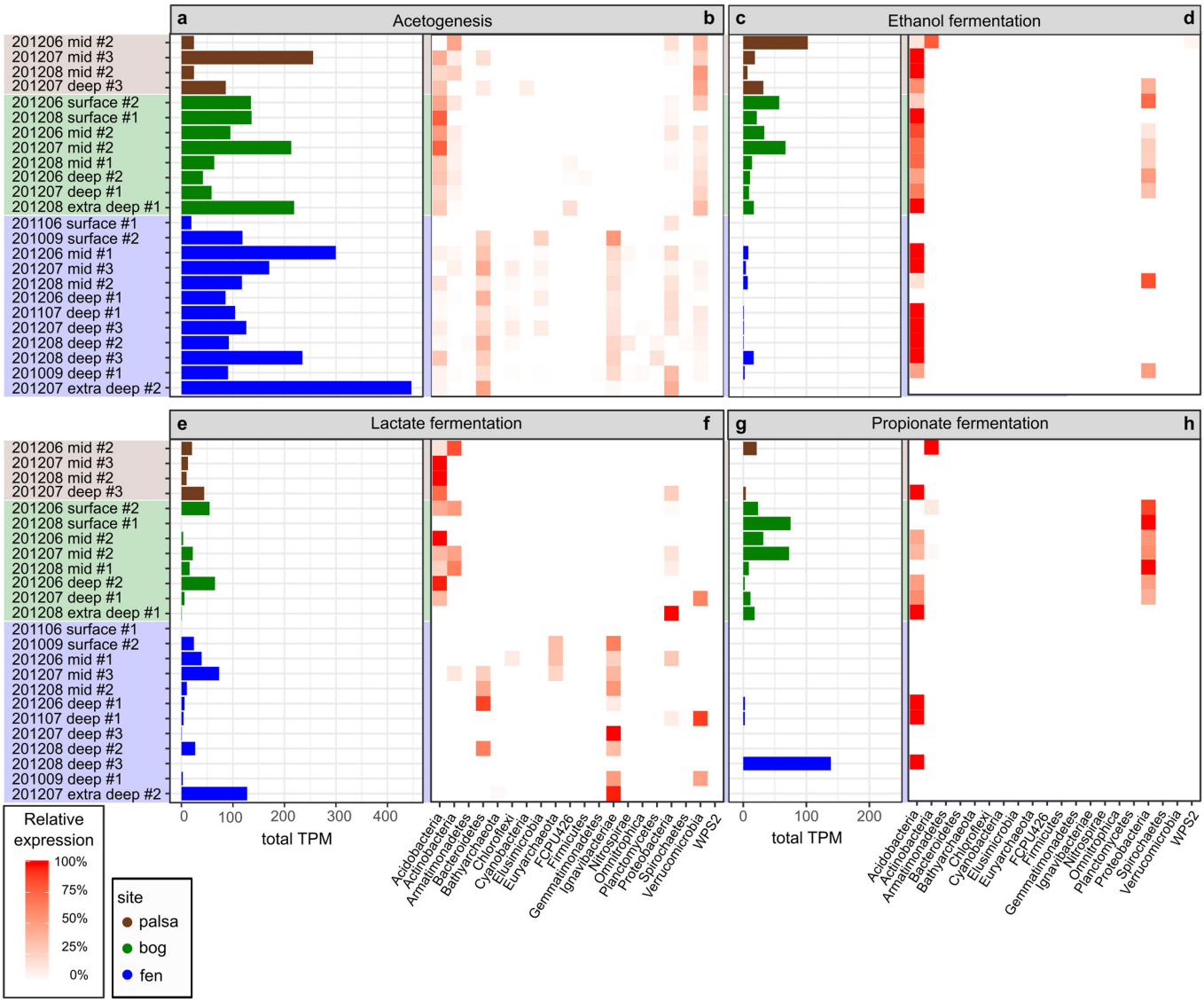
Extended Data Fig. 5 | Monosaccharide degradation pathway prevalence at Stordalen Mire. a, As in Fig. 2, 97% dereplicated MAGs are shown as circles ('MAG abundance'), where the radius of the circle represents the average relative abundance of that genome in the palsa, bog or fen. **b**, As in Fig. 2, the total relative abundance of genomes encoding the pathway is shown among the entire community. Sample numbers: $n = 53$, 65 and 70 biologically independent samples for palsa, bog and fen, respectively.

© 2018 Springer Nature Limited. All rights reserved.



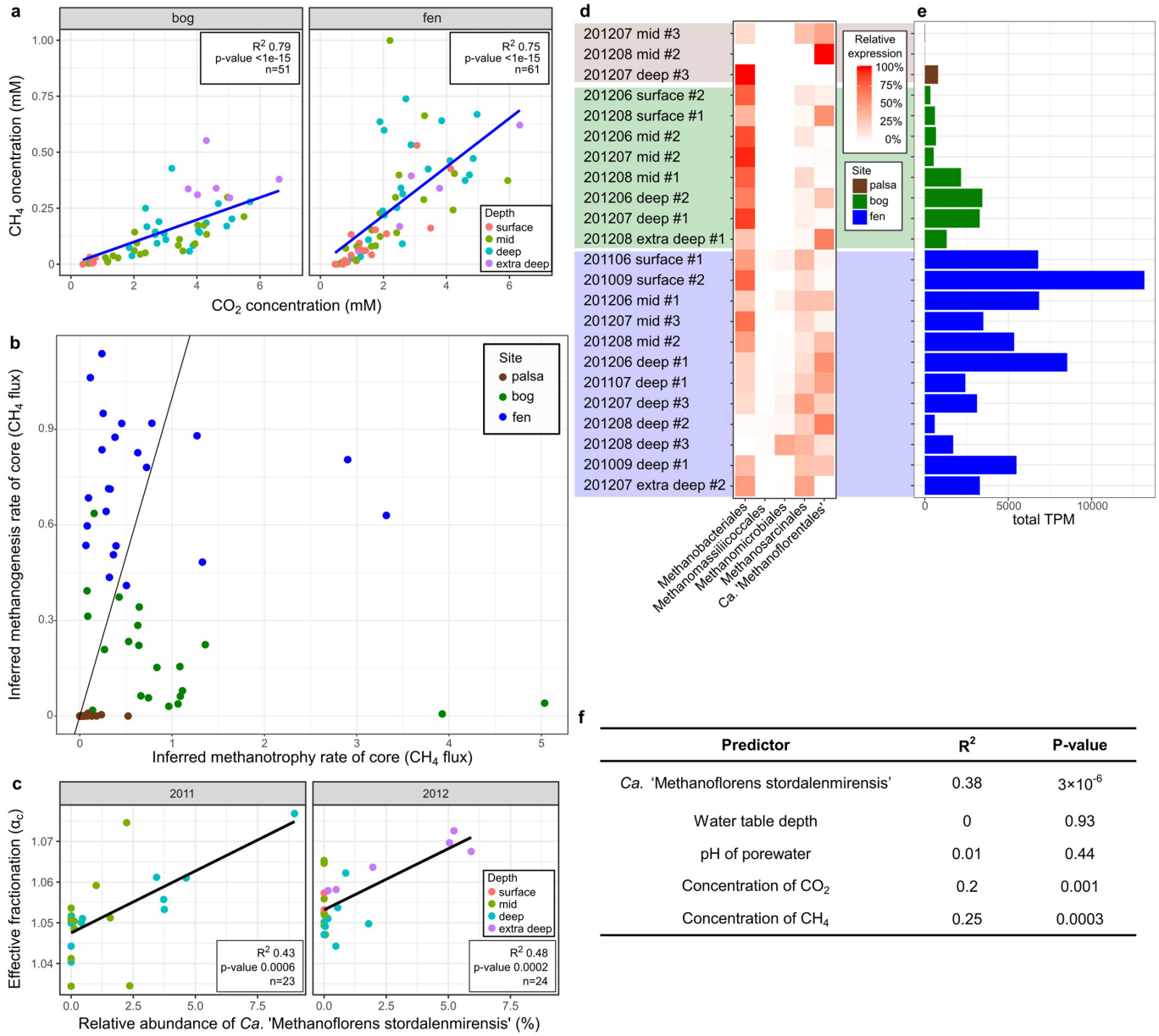
Extended Data Fig. 6 | Xylose degradation pathways at Stordalen Mire. **a**, Diagram of xylose degradation pathways. **b**, Venn diagram showing how each xylose breakdown pathway is shared among the Stordalen Mire MAGs. Percentages represent the proportion compared to all Stordalen genomes encoding a xylose degradation pathway. In the metaproteomes, genomes Acidobacteria_bog_390, Actinobacteria_fen_455 and Actinobacteria_bog_808 expressed a protein specific to oxidoreductase pathways and a protein specific to the isomerase pathway. In the metatranscriptomes, Acidobacteria_palsa_248, Acidobacteria_bog_370, Acidobacteria_bog_390, Actinobacteria_fen_455, Actinobacteria_bog_586, Actinobacteria_bog_808 and Planctomycetes_fen_1346 expressed a protein specific to oxidoreductase pathways and a protein specific to the isomerase pathway. **c**–**h**, Gene expression of xylose degradation pathways. Average expression of genes in the canonical bacterial xylose isomerase (**c**, **d**), oxidoreductase (**e**, **f**) and xylanate dehydratase pathways (**g**, **h**) are depicted across the thaw gradient. Samples analysed with metatranscriptomics are described by the date of sampling, core number and depth. **c**, **e**, **g**, Relative contribution of each phylum to the total TPM of the enzyme class observed in the metatranscriptomes. **d**, **f**, **h**, Total TPM of all expressed genes in the sample.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Gene expression of fermentation pathways. Samples analysed with metatranscriptomics are described by the date of sampling, core number and depth. **a**, **c**, **e**, **g**, Total TPM of each fermentation pathway in the metatranscriptomes. **b**, **d**, **f**, **h**, Relative contribution of each phylum to the total TPM of each pathway.

© 2018 Springer Nature Limited. All rights reserved.

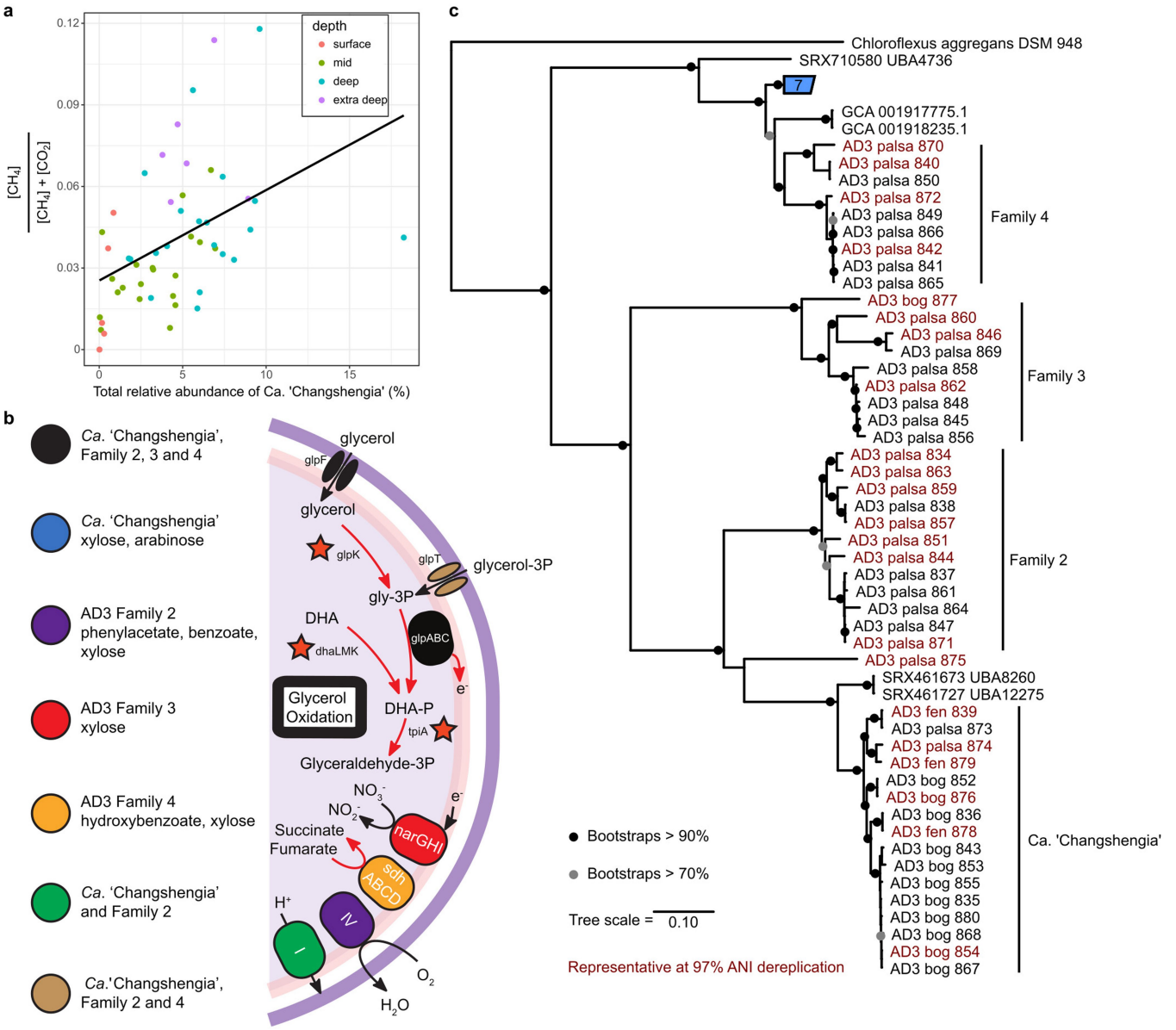


f

Predictor	R^2	P-value
Ca. 'Methanoflorens stordalenmirensis'	0.38	3×10^{-6}
Water table depth	0	0.93
pH of porewater	0.01	0.44
Concentration of CO_2	0.2	0.001
Concentration of CH_4	0.25	0.0003

Extended Data Fig. 8 | Correlation of microbial and geochemical data. **a**, CO_2 and CH_4 concentrations in porewater derived from the bog and fen. The blue line shown is a line of best fit, forced through the origin. Dots indicate the samples, with colours indicating the sample depth. The concentrations are correlated, and the CH_4 concentrations are much lower than the CO_2 concentrations in both sites. Sample numbers: $n = 51$ (bog) and 61 (fen) biologically independent samples. **b**, Methanogenesis versus methanotrophy rates. Each point represents the average relative abundance of methanotrophs and methanogens across all samples in a single core, multiplied by the rate of methane generation or consumption inferred from previous culture-based measurements (2.345 and 20.1 fmol CH_4 h^{-1} per cell of methanogenesis and methanotrophy, respectively, see Methods). The line represents the 1:1 ratio. Inferred fluxes were calculated using relative abundance of methanogenic or methanotrophic lineages so rates are only intended for comparison between the *x* and *y* axes, rather than as an absolute measure of CH_4 flux. Methanotrophy appears to mitigate a significant proportion of the CH_4 generated in the bog sites. **c**, Correlation of the relative abundance of *Ca.* 'Methanoflorens stordalenmirensis' with the isotopic fractionation of methane (α_C) dissolved in paired porewater samples taken from the bog. Previously observed in 2011 using 16S rRNA gene amplicon sequencing[33], the correlation is confirmed here using genome-centric metagenomic techniques on the 2011 samples, as well as in a new year of sampling in 2012. Sample numbers: $n = 23$ (2011) and 24 (2012) biologically independent samples. **d**, **e**, Expression of methanogenesis marker gene *mcrA* across the thaw gradient. Samples analysed with metatranscriptomics are described by the date of sampling, core number and depth. **d**, Relative contribution of each methanogenic order to the total TPM. **e**, Relative contribution of all *mcrA* genes in the metatranscriptome. Metaproteomes revealed the expression of 289 hydrogenotrophic McrA proteins across 13 samples, as well as 78 acetoclastic McrA proteins across eight samples (Supplementary Data 2). **f**, Linear regression analysis for predicting effective fractionation (α_c) of CH_4 from environmental variables and *Ca.* 'Methanoflorens stordalenmirensis' abundances in the bog. *Ca.* 'Methanoflorens stordalenmirensis' abundance exceeds bulk geochemical parameters in predicting the effective fractionation of CH_4. Each line is the result of a linear regression of the specified measurement against the α_c of CH_4 in bog porewater samples taken in 2011 and 2012 ($n = 47$ biologically independent samples).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Candidate phylum Dormibacteraeota (AD3) genus *Ca.* 'Changshengia' at Stordalen Mire. a, Total relative abundance of the genus *Ca.* 'Changshengia' correlated with the fraction of the concentration of C mineralized to CO_2 versus CH_4 in the bog porewater samples ($R^2 = 0.19$, $P = 0.001$, $n = 51$ biologically independent samples). Each point represents an individual sample from 2012, with its colour representing the depth in the core from which the sample was taken. **b**, Metabolic reconstruction of genomes belonging to the candidate phylum AD3 genus *Ca.* 'Changshengia' correlating with the CH_4:CO_2 concentration ratio in porewater from 2012 bog samples. Genomes from four clades within the AD3 were assembled from across Stordalen Mire. Enzyme colour indicates the families that share that metabolic potential, as outlined in the legend on the left. Arrow colouring indicates whether expression was detected (red arrows) or not detected (black arrows) for genes encoding the enzyme in any of the 24 metatranscriptomes. Orange stars indicate detection of protein expression in any of the 22 metaproteomes from the *Ca.* 'Changshengia' and related genomes. All four lineages encode the potential to oxidize glycerol anaerobically through glycerol transporter (*glpF*), glycerol kinase (*glpK*) and a membrane-bound glycerol-3-phosphate dehydrogenase (*glpABC*), entering glycolysis via dihydroxyacetone phosphate processed to glyceraldehyde-3-phosphate by the triosephosphate isomerase (*tpiA*). Other glycerol derivatives such as glycerol-3-phosphate could be imported (*glpT*) by this and other family members, and dihydroxyacetone phosphate can also be processed using the PTS-dependent dihydroxyacetone kinase (*dhaLMK*) complex. Sinks for the electrons generated from the oxidation of glycerol also varied between the different lineages, with *Ca.* 'Changshengia' and clade 1 having a H^+-translocating complex I NADH:oxidoreductase, while clade 1 also has a high affinity cytochrome oxidase complex IV, clade 2 genomes encode only a nitrate reductase (*narGHI*) and clade 4 genomes only a fumarate reductase (*sdhABCD*). These differences are likely to lead to the differentiation of the niches that each lineage occupies across different sites and depths of the mire. Lineages were considered positive for genes or complexes based on the presence of sequences with 80% homology in 50% of the genomes. **c**, Phylogenetic subtree showing the family groupings of AD3 for the metabolic analysis. Representative genomes from the 97% average nucleotide identity (ANI) dereplication are indicated in red. Bootstrap support is indicated at the nodes for values over 70% or 90% in grey and black, respectively. Blue clade indicates cluster of seven UBA and RefSeq genomes.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Genomes with high prevalence in specific sites and depths

Genome	Environment	Num present	Total samples	Prevalence (%)
Acidobacteria_palsa_348	palsa surface	18	18	100
Acidobacteria_bog_233	bog surface	18	19	95
Gammaproteobacteria_bog_1200	bog surface	18	19	95
Actinobacteria_bog_635	bog surface	19	19	100
Actinobacteria_bog_723	bog surface	18	19	95
Actinobacteria_bog_766	bog surface	18	19	95
Acidobacteria_bog_226	bog deep	19	20	95
Acidobacteria_bog_406	bog deep	19	20	95
Acidobacteria_fen_408	bog deep	20	20	100
Actinobacteria_bog_808	bog deep	19	20	95
Acidobacteria_bog_218	bog extra deep	6	6	100
Acidobacteria_bog_226	bog extra deep	6	6	100
Acidobacteria_bog_406	bog extra deep	6	6	100
Acidobacteria_fen_408	bog extra deep	6	6	100
Deltaproteobacteria_fen_1087	fen surface	22	23	96
Euryarchaeota_fen_64	fen surface	21	23	91
Deltaproteobacteria_fen_1087	fen mid	18	18	100
Deltaproteobacteria_fen_1087	fen deep	21	23	91
Euryarchaeota_fen_64	fen deep	21	23	91
Deltaproteobacteria_fen_1087	fen extra deep	6	6	100

Genomes shown are representative genomes from the 97% dereplicated set that are present at >1% relative abundance in >90% of samples from the site and depth shown in the 'Environment' column. The 'Num present' column indicates the number of samples in which it is found and 'Total samples' indicates the total number of samples available for that environment.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Overview of proteins detected using metaproteomics

Sample	# of spectra	Top KOs	Top phyla
201206_P123M	31	K02650: pilA; type IV pilus assembly protein PilA, K04077: groEL, HSPD1; chaperonin GroEL, K14028: mdh1, mxaF; methanol dehydrogenase (cytochrome c) subunit 1	Acidobacteria (71%), AD3 (19.4%), Proteobacteria (6.5%), Euryarchaeota (3.2%)
201206_S123D	1687	K00024: mdh; malate dehydrogenase, K00320: mer; 5,10-methylenetetrahydromethanopterin reductase, K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K02358: tuf, TUFM; elongation factor Tu, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04043: dnaK, HSPA9; molecular chaperone DnaK, K04077: groEL, HSPD1; chaperonin GroEL, K05349: bglX; beta-glucosidase	Acidobacteria (56.4%), AD3 (21.6%), Euryarchaeota (13.9%), Proteobacteria (3.4%), Actinobacteria (3.2%), Verrucomicrobia (0.5%), Ignavibacteriae (0.3%), Thaumarchaeota (0.2%), WPS2 (0.2%), Bacteroidetes (0.1%)
201206_S123M	700	K00024: mdh; malate dehydrogenase, K00138: aldB; aldehyde dehydrogenase, K02358: tuf, TUFM; elongation factor Tu, K03286: TC.OOP; OmpA-OmpF porin, OOP family, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04043: dnaK, HSPA9; molecular chaperone DnaK, K04077: groEL, HSPD1; chaperonin GroEL, K05349: bglX; beta-glucosidase, K14028: mdh1, mxaF; methanol dehydrogenase (cytochrome c) subunit 1, K16087: TC.FEV.OM3, tbpA, hemR, lbpA, hpuB, bhuR, hugA, hmbR; hemoglobin/transferrin/lactoferrin receptor protein	Acidobacteria (76.1%), AD3 (14.9%), Proteobacteria (3.9%), Euryarchaeota (3.4%), Actinobacteria (1.4%), Ignavibacteriae (0.1%), Verrucomicrobia (0.1%)
201207_E2X	101	K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K00428: E1.11.1.5; cytochrome c peroxidase, K00864: glpK, GK; glycerol kinase, K01114: plc; phospholipase C, K03286: TC.OOP; OmpA-OmpF porin, OOP family, K11254: H4; histone H4, K12340: tolC; outer membrane protein	Acidobacteria (45.5%), Euryarchaeota (32.7%), AD3 (17.8%), Proteobacteria (2%), Gemmatimonadetes (1%), Ignavibacteriae (1%)
201207_E3D	153	K00193: cdhC; acetyl-CoA decarbonylase/synthase complex subunit beta, K00320: mer; 5,10-methylenetetrahydromethanopterin reductase, K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K00864: glpK, GK; glycerol kinase, K01895: ACSS, acs; acetyl-CoA synthetase, K02040: pstS; phosphate transport system substrate-binding protein, K02731: PSMA7; 20S proteasome subunit alpha 4, K03622: ssh10b; archaea-specific DNA-binding protein	Euryarchaeota (73.9%), Acidobacteria (17%), AD3 (7.8%), Actinobacteria (0.7%), Proteobacteria (0.7%)
201207_E3M	537	K00024: mdh; malate dehydrogenase, K00193: cdhC; acetyl-CoA decarbonylase/synthase complex subunit beta, K00320: mer; 5,10-methylenetetrahydromethanopterin reductase, K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K00864: glpK, GK; glycerol kinase, K01895: ACSS, acs; acetyl-CoA synthetase, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04077: groEL, HSPD1; chaperonin GroEL	Euryarchaeota (45.3%), Acidobacteria (26.6%), AD3 (17.3%), Proteobacteria (5.4%), Chloroflexi (1.3%), Ignavibacteriae (1.1%), Verrucomicrobia (1.1%), Nitrospirae (0.7%), Actinobacteria (0.6%), Bacteroidetes (0.4%)
201207_P3D	321	K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K00406: ccoP; cytochrome c oxidase cbb3-type subunit III, K00428: E1.11.1.5; cytochrome c peroxidase, K02650: pilA; type IV pilus assembly protein PilA, K03286: TC.OOP; OmpA-OmpF porin, OOP family, K04077: groEL, HSPD1; chaperonin GroEL, K10696: BRE1; E3 ubiquitin-protein ligase BRE1, K16087: TC.FEV.OM3, tbpA, hemR, lbpA, hpuB, bhuR, hugA, hmbR; hemoglobin/transferrin/lactoferrin receptor protein, K16089: TC.FEV.OM2, cirA, cfrA, hmuR; outer membrane receptor for ferrienterochelin and colicins, K16090: fiu; catecholate siderophore receptor	Acidobacteria (80.7%), AD3 (14%), Euryarchaeota (4.7%), Proteobacteria (0.6%)
201207_P3M	78	K00600: glyA, SHMT; glycine hydroxymethyltransferase, K02358: tuf, TUFM; elongation factor Tu, K02601: nusG; transcriptional antiterminator NusG, K03046: rpoC; DNA-directed RNA polymerase subunit beta', K03704: cspA; cold shock protein (beta-ribbon, CspA family), K04043: dnaK, HSPA9; molecular chaperone DnaK, K04077: groEL, HSPD1; chaperonin GroEL, K06006: cpxP; periplasmic protein CpxP, K14028: mdh1, mxaF; methanol dehydrogenase (cytochrome c) subunit 1, K17734: aprX; serine protease AprX	Acidobacteria (64.1%), AD3 (29.5%), Euryarchaeota (3.8%), Proteobacteria (2.6%)
201207_S1D	1563	K00024: mdh; malate dehydrogenase, K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K02358: tuf, TUFM; elongation factor Tu, K03286: TC.OOP; OmpA-OmpF porin, OOP family, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04043: dnaK, HSPA9; molecular chaperone DnaK, K04077: groEL, HSPD1; chaperonin GroEL, K05349: bglX; beta-glucosidase	Acidobacteria (58.5%), Euryarchaeota (17.1%), AD3 (16.9%), Proteobacteria (3.2%), Actinobacteria (3%), Verrucomicrobia (0.6%), Ignavibacteriae (0.3%), Bacteroidetes (0.2%), FCPU426 (0.1%), Chloroflexi (0.1%)
201207_S2M	797	K00024: mdh; malate dehydrogenase, K01114: plc; phospholipase C, K02358: tuf, TUFM; elongation factor Tu, K03286: TC.OOP; OmpA-OmpF porin, OOP family, K03695: clpB; ATP-dependent Clp protease ATP-binding subunit ClpB, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04043: dnaK, HSPA9; molecular chaperone DnaK, K04077: groEL, HSPD1; chaperonin GroEL, K04749: rsbV; anti-sigma B factor antagonist, K14028: mdh1, mxaF; methanol dehydrogenase (cytochrome c) subunit 1	Acidobacteria (78.3%), AD3 (13%), Proteobacteria (5%), Euryarchaeota (2%), Actinobacteria (1.1%), Ignavibacteriae (0.3%), Chloroflexi (0.1%), Verrucomicrobia (0.1%)
201208_E2D	725	K00319: mtd; methylenetetrahydromethanopterin dehydrogenase, K00320: mer; 5,10-methylenetetrahydromethanopterin reductase, K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K00864: glpK, GK; glycerol kinase, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04077: groEL, HSPD1; chaperonin GroEL, K09495: CCT3, TRIC5; T-complex protein 1 subunit gamma, K16087: TC.FEV.OM3, tbpA, hemR, lbpA, hpuB, bhuR, hugA, hmbR; hemoglobin/transferrin/lactoferrin receptor protein	Euryarchaeota (50.5%), Acidobacteria (27.4%), AD3 (17.2%), Proteobacteria (1.5%), Chloroflexi (1.2%), Actinobacteria (0.6%), Ignavibacteriae (0.4%), Verrucomicrobia (0.4%), Fibrobacteres (0.4%), Planctomycetes (0.1%)
201208_E2M	789	K00193: cdhC; acetyl-CoA decarbonylase/synthase complex subunit beta, K00320: mer; 5,10-methylenetetrahydromethanopterin reductase, K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K00864: glpK, GK; glycerol kinase, K01895: ACSS, acs; acetyl-CoA synthetase, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04077: groEL, HSPD1; chaperonin GroEL, K14028: mdh1, mxaF; methanol dehydrogenase (cytochrome c) subunit 1	Euryarchaeota (57.5%), Acidobacteria (22.2%), AD3 (13.1%), Proteobacteria (4.1%), Actinobacteria (1%), Verrucomicrobia (0.5%), Ignavibacteriae (0.4%), Chloroflexi (0.4%), Nitrospirae (0.3%), Bacteroidetes (0.3%)
201208_E3D	108	K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K03046: rpoC; DNA-directed RNA polymerase subunit beta', K03286: TC.OOP; OmpA-OmpF porin, OOP family, K04077: groEL, HSPD1; chaperonin GroEL, K06882: uncharacterized protein, K09495: CCT3, TRIC5; T-complex protein 1 subunit gamma, K14028: mdh1, mxaF; methanol dehydrogenase (cytochrome c) subunit 1, K16087: TC.FEV.OM3, tbpA, hemR, lbpA, hpuB, bhuR, hugA, hmbR; hemoglobin/transferrin/lactoferrin receptor protein	Acidobacteria (50%), Euryarchaeota (25.9%), AD3 (18.5%), Proteobacteria (3.7%), Actinobacteria (0.9%), Fibrobacteres (0.9%)
201208_P2M	160	K00024: mdh; malate dehydrogenase, K01114: plc; phospholipase C, K02338: dnaN; DNA polymerase III subunit beta, K03046: rpoC; DNA-directed RNA polymerase subunit beta', K03286: TC.OOP; OmpA-OmpF porin, OOP family, K04043: dnaK, HSPA9; molecular chaperone DnaK, K04077: groEL, HSPD1; chaperonin GroEL, K06006: cpxP; periplasmic protein CpxP, K14028: mdh1, mxaF; methanol dehydrogenase (cytochrome c) subunit 1, K17734: aprX; serine protease AprX	Acidobacteria (80%), AD3 (15.6%), Proteobacteria (1.9%), Euryarchaeota (1.2%), Chloroflexi (0.6%), Actinobacteria (0.6%)
201208_S1M	2034	K00024: mdh; malate dehydrogenase, K00320: mer; 5,10-methylenetetrahydromethanopterin reductase, K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K01114: plc; phospholipase C, K02358: tuf, TUFM; elongation factor Tu, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04043: dnaK, HSPA9; molecular chaperone DnaK, K04077: groEL, HSPD1; chaperonin GroEL	Acidobacteria (62.8%), AD3 (16.7%), Euryarchaeota (14.2%), Proteobacteria (3.2%), Actinobacteria (2%), Ignavibacteriae (0.2%), Verrucomicrobia (0.2%), Thaumarchaeota (0.2%), Bathyarchaeota (0.1%), Bacteroidetes (0.1%)
201208_S1X	3951	K00024: mdh; malate dehydrogenase, K00320: mer; 5,10-methylenetetrahydromethanopterin reductase, K00399: mcrA; methyl-coenzyme M reductase alpha subunit, K00401: mcrB; methyl-coenzyme M reductase beta subunit, K00402: mcrG; methyl-coenzyme M reductase gamma subunit, K02358: tuf, TUFM; elongation factor Tu, K03286: TC.OOP; OmpA-OmpF porin, OOP family, K03737: por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase, K04043: dnaK, HSPA9; molecular chaperone DnaK, K04077: groEL, HSPD1; chaperonin GroEL	Acidobacteria (44.3%), Euryarchaeota (33.7%), AD3 (16.4%), Actinobacteria (2.5%), Proteobacteria (1.7%), Verrucomicrobia (0.4%), Ignavibacteriae (0.2%), Chloroflexi (0.2%), Firmicutes (0.2%), FCPU426 (0.1%)

The third column shows the ten most abundant KEGG Orthology (KO) groups detected, where the total spectral count for that group was two or more. The fourth column shows the relative abundance of spectral counts from each phylum.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0342-5

Leukaemia hijacks a neural mechanism to invade the central nervous system

Hisayuki Yao[1,6,9], Trevor T. Price[1,9], Gaia Cantelli[1], Brandon Ngo[1], Matthew J. Warner[1,7], Lindsey Olivere[1], Sarah M. Ridge[1], Elizabeth M. Jablonski[2,8], Joseph Therrien[3], Stacey Tannheimer[3], Chad M. McCall[4], Anjen Chenn[5] & Dorothy A. Sipkins[1]*

Acute lymphoblastic leukaemia (ALL) has a marked propensity to metastasize to the central nervous system (CNS). In contrast to brain metastases from solid tumours, metastases of ALL seldom involve the parenchyma but are isolated to the leptomeninges, which is an infrequent site for carcinomatous invasion. Although metastasis to the CNS occurs across all subtypes of ALL, a unifying mechanism for invasion has not yet been determined. Here we show that ALL cells in the circulation are unable to breach the blood–brain barrier in mice; instead, they migrate into the CNS along vessels that pass directly between vertebral or calvarial bone marrow and the subarachnoid space. The basement membrane of these bridging vessels is enriched in laminin, which is known to coordinate pathfinding of neuronal progenitor cells in the CNS. The laminin receptor α6 integrin is expressed in most cases of ALL. We found that α6 integrin–laminin interactions mediated the migration of ALL cells towards the cerebrospinal fluid in vitro. Mice with ALL xenografts were treated with either a PI3Kδ inhibitor, which decreased α6 integrin expression on ALL cells, or specific α6 integrin–neutralizing antibodies and showed significant reductions in ALL transit along bridging vessels, blast counts in the cerebrospinal fluid and CNS disease symptoms despite minimally decreased bone marrow disease burden. Our data suggest that α6 integrin expression, which is common in ALL, allows cells to use neural migratory pathways to invade the CNS.

In the absence of CNS-directed prophylactic treatment, involvement of the CNS during disease progression occurs in 30–70% of patients with ALL[1]. Metastasis to the CNS is observed in all subtypes of ALL, suggesting that lymphoblasts utilize a conserved molecular mechanism to invade the CNS. ALL relapse in the CNS predicts poor outcomes, but treatment options remain limited[2].

The enzyme PI3K has an important role transducing extracellular signals that regulate cell growth and survival[3]. The delta isoform of PI3K (PI3Kδ) is uniquely expressed in immune cells and neurons, and the PI3Kδ inhibitor idelalisib is approved to treat indolent B cell lymphomas[4,5]. The efficacy of idelalisib in acute B cell malignancies has not yet been established. Therefore, we examined the role of PI3Kδ inhibition in the Nalm-6 model of ALL. The Nalm-6 pre-B ALL cell line was derived from a patient who suffered a CNS disease relapse. It generates a reproducible pattern of disease in severe combined immunodeficiency (SCID) mice that mimics ALL development in patients, including CNS invasion. At approximately 40 days after intravenous engraftment of Nalm-6 cells, all mice show symptoms of CNS involvement, which occur before death from progressive bone marrow disease, and lower extremity paresis is the usual clinical end point for euthanasia.

PI3Kδ blockade reduces CNS disease burden

Using a PI3Kδ inhibitor tool compound, GS-649443, that has a suitable pharmacokinetic profile in mice, we treated Nalm-6-engrafted mice daily with vehicle or inhibitor beginning on day 1 or day 20 after engraftment (Fig. 1a) and continuing until the clinical end point for euthanasia was reached. We observed an almost 50% prolongation of survival in early and late GS-649443 treatment arms, with most GS-649443-treated mice euthanized secondary to bone marrow failure (Fig. 1b). Notably, the incidence of CNS disease symptoms at time of euthanasia in treated mice was decreased by approximately six- and threefold when comparing control and drug treatment in the early and late treatment-initiation groups, respectively (Fig. 1c).

To determine whether the low incidence of CNS involvement in these mice reflected delayed disease progression in the periphery, we compared the tumour burden in bone marrow, spleen and CNS of GS-649443- or vehicle-treated mice at matched time points. Paired PI3Kδ inhibitor- and vehicle-treated mice were euthanized simultaneously when either reached a clinical end point. There was no significant difference between bone marrow or splenic Nalm-6 disease burden or peripheral blood cell counts in inhibitor- compared to vehicle-treated mice (Fig. 1d, e and Extended Data Fig. 1a, b). By contrast, there was an approximately 50% decrease in CNS disease burden among paired mice (Fig. 1e and Extended Data Fig. 1c).

To confirm that our results were not isolated to the Nalm-6 cell line, we next tested the efficacy of PI3Kδ inhibition in the RCH-ACV ALL model, which exhibits symptomatic CNS involvement with similar frequency, and in non-obese diabetic/SCID/IL-2Rγ mice engrafted with primary human ALL cells. Similar to our results in Nalm-6-engrafted mice, single-agent PI3Kδ inhibition markedly decreased CNS disease in these models, but only minimally diminished peripheral disease burden (Fig. 1f, g and Extended Data Fig. 1d–i). These findings suggested that GS-649443 inhibited progression of ALL in the CNS independently of its effects on peripheral disease. Initially, we hypothesized that PI3Kδ signalling was necessary for ALL proliferation or survival in the CNS microenvironment. We therefore first examined whether the PI3Kδ inhibitor GS-649443 could penetrate the blood–brain barrier in healthy or leukaemic mice. Whereas mean serum concentrations of GS-649443 were approximately 200–350 nM (Fig. 2a and Extended Data Fig. 2a), compound concentrations in the brain tissue of healthy or end-stage leukaemic mice were less than 5 nM, well below inhibitory concentrations (Extended Data Fig. 2c, d). Analysis of blasts in the cerebrospinal

[1]Department of Medicine, Division of Hematologic Malignancies and Cellular Therapy, Duke University, Durham, NC, USA. [2]Department of Medicine, Division of Hematology and Oncology, University of Chicago, Chicago, IL, USA. [3]Gilead Sciences, Inc., Foster City, CA, USA. [4]Department of Pathology, Duke University, Durham, NC, USA. [5]LabCorp, Research Triangle Park, NC, USA. [6]Present address: Department of Hematology, The National Hospital Organization Hiroshima-Nishi Medical Center, Hiroshima-ken, Japan. [7]Present address: Vitrisa Therapeutics, Inc., Durham, NC, USA. [8]Present address: The Graduate School, Northwestern University, Evanston, IL, USA. [9]These authors contributed equally: Hisayuki Yao, Trevor T. Price.. *e-mail: dorothy.sipkins@duke.edu

© 2018 Springer Nature Limited. All rights reserved.

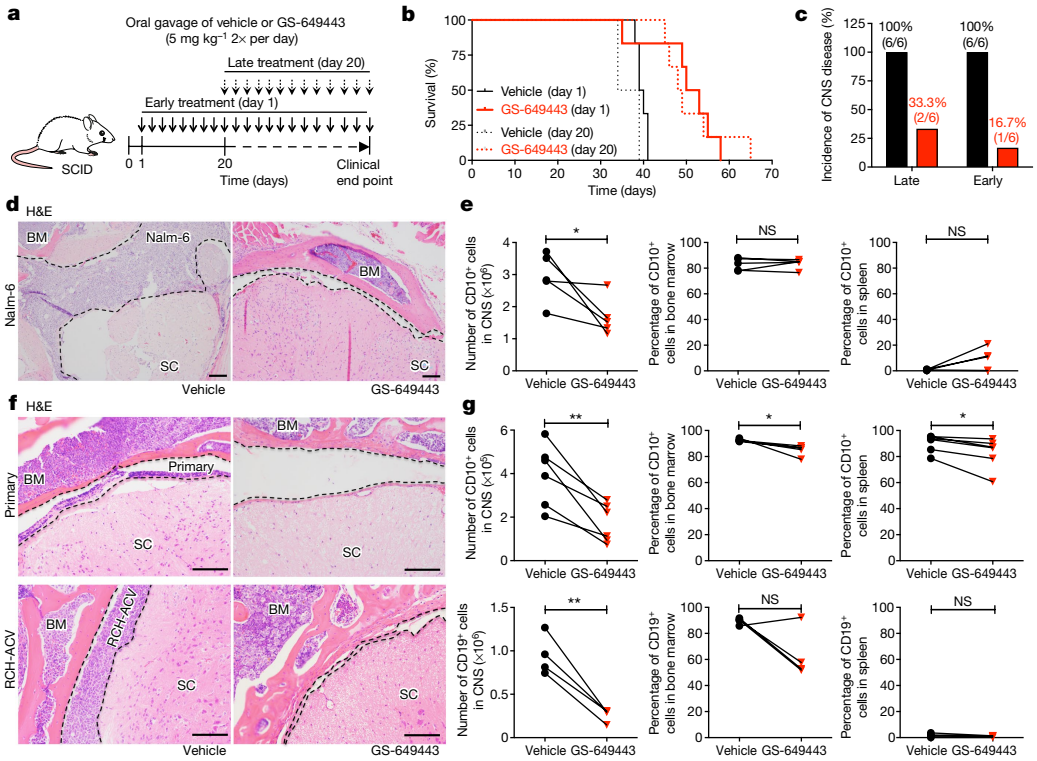


Fig. 1 | PI3Kδ inhibition blocks progression of ALL in the CNS in vivo. **a**, Schematic of treatment of mice with GS-649443. **b**, Kaplan–Meier survival curves. Two-sided log rank Mantel-Cox test, $n = 6$ mice per treatment group, late treatment: $P = 0.0039$, early treatment: $P = 0.0146$. **c**, Incidence of hindlimb paralysis at time of euthanasia. $n = 6$ mice per treatment group. **d**, Nalm-6 ALL cells in subarachnoid space (dashed line) of the spinal cord (SC). Representative sections of the vertebrae are shown. H&E, haematoxylin and eosin. $n = 14$ GS-649443-treated mice, 7 vehicle-treated mice. **e**, Disease burden (number of $CD10^+$ ALL cells) in the CNS, bone marrow and spleen of vehicle- and GS-649443-treated Nalm-6-engrafted mice euthanized at matched time points. Paired two-sided Student's *t*-test, $n = 5$ mice per treatment group, $P = 0.0473$. **f**, Primary (1°) human and RCH-ACV ALL in subarachnoid space of spinal cord. Representative sections of the vertebrae are shown. Primary ALL: $n = 6$ GS-649443-treated mice, 7 vehicle-treated mice; RCH-ACV: $n = 4$ mice per treatment group. **g**, Disease burden (indicated by $CD10^+$ or $CD19^+$ ALL cells) in the CNS, bone marrow and spleen of vehicle- or GS-649443-treated mice engrafted with primary human ALL and RCH-ACV ALL cells. Paired two-sided Student's *t*-test, primary human cells: $n = 5$ mice per treatment group; RCH-ACV: $n = 4$ mice per treatment group. $^*P = < 0.05$, $^{**}P < 0.01$. Scale bars, 100 μm.

fluid (CSF) isolated from GS-649443- and vehicle-treated mice revealed no differences in apoptotic or mitotic indices, further indicating that the pronounced effect of PI3Kδ inhibition on CNS disease progression did not occur through direct effects on ALL growth or viability (Extended Data Fig. 2e–g). We therefore turned our attention to the potential function of PI3Kδ in ALL metastasis from the periphery to the CNS.

PI3K modulates the motility of ALL cells

In addition to its roles in cell growth, PI3K is a key mediator of migration, controlling actomyosin contractility through the downstream Rho and FAK pathways[6–8]. We therefore evaluated the role of PI3Kδ in ALL cell motility in vitro. PI3Kδ inhibition with either GS-649443 or idelalisib decreased chemotaxis of ALL cell lines and primary ALL cells in transwell migration assays (Fig. 2b and Extended Data Fig. 2b). ALL migration was independent of both AKT and ROCK, yet highly dependent on MLCK activity, suggesting that the FAK signalling pathway downstream of PI3K mediated the inhibition of ALL migration[9] (Fig. 2b and Extended Data Fig. 2b, h, i). Western blot analysis indicated that myosin light chain (MLC) activity in GS-649443-treated ALL cells was reduced (Extended Data Fig. 3a), leading us to hypothesize that PI3Kδ inhibition in vivo blocked the development of CNS disease by broadly paralysing the motility of ALL cells. In previously published work, we have shown that ALL and other tumour cells metastasize to the bone marrow through sinusoidal vasculature located in specific anatomic regions[10]. Proliferating cells migrate away from these sinusoidal vessels over time[11,12]. We hypothesized that if ALL cell motility was broadly inhibited by PI3Kδ blockade, then proliferating Nalm-6 cells in the bone marrow of treated mice would fail to migrate from the sites of initial engraftment. We used intravital microscopy to track the location of ALL cells in the bone marrow at serial time points after engraftment in mice, but found no significant differences in intra-bone marrow migration in treated mice compared to mice treated with vehicle control (Fig. 2c and Extended Data Fig. 3b). These data suggested that ALL cells had not lost all migratory potential in vivo as a consequence of PI3K inhibition, but that ALL migration into the CNS was regulated by a specific mechanism(s) upstream of MLCK. To identify candidate molecules, we performed a microarray analysis of blasts isolated from the bone marrow and CSF of Nalm-6-engrafted mice. Genes in both focal adhesion and contractility signalling pathways were strongly downregulated in the bone marrow and CSF blasts of GS-649443-treated mice (Extended Data Fig. 4a–c). Among multiple genes within the focal adhesion pathway was *ITGA6*, which is intriguing because of its dual relationship during cell migration and CNS development[13–15].

PI3K regulates α6 integrin in ALL cells

ITGA6 encodes the α6 integrin, which dimerizes with β1 or β4 integrin to form receptors that specifically bind to the extracellular matrix (ECM) molecule laminin[16]. Notably, *Itga6* knockout mice display developmental defects in cerebral cortical organization, with abnormal neurite outgrowths studding the surface of the brain[14]. In vitro, neural stem/precursor cells (NSPCs) have been shown to migrate selectively along laminin matrices in an α6 integrin-dependent manner[15]. α6 integrin has not been extensively studied in leukaemia, but analysis of samples obtained from patients with primary ALL have demonstrated that it is expressed by the majority of B and T ALL and that its expression is persistent or intensified in residual disease following chemotherapy[17–19].

© 2018 Springer Nature Limited. All rights reserved.

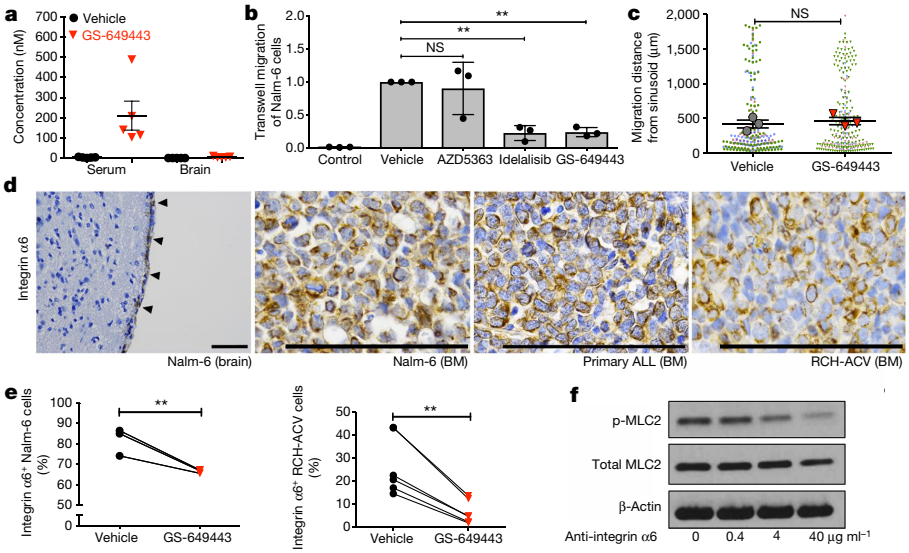


Fig. 2 | PI3Kδ regulates aspects of ALL cell motility and expression of the laminin receptor α6 integrin. a, Serum and brain tissue concentrations of GS-649443 in leukaemic mice. Data are mean ± s.e.m.; $n = 5$ mice per group. **b**, Effect of PI3Kδ and Akt inhibition on transwell migration of ALL. Data are mean ± s.e.m.; ANOVA with Tukey test, $n = 3$ biologically independent experiments; $^{*}P < 0.05$, $^{**}P < 0.01$. **c**, Nalm-6 migration in calvarial bone marrow over 10 days in vehicle- versus GS-649443-treated mice. Data are mean of the sample means ± s.e.m.; $n = 3$ mice per treatment group, data points from individual mice distinguished by colour, paired two-sided Student's *t*-test. **d**, α6 integrin expression of ALL cells in brain and bone marrow (BM) in vivo. Representative immunohistochemistry sections are shown. Arrowheads indicate ALL cells in meninges. Nalm-6 brain: $n = 7$ mice, Nalm-6 bone marrow: $n = 12$ mice, primary ALL bone marrow: $n = 4$ mice, RCH-ACV bone marrow: $n = 11$ mice. **e**, Flow cytometry of ALL α6 integrin expression following in vitro treatment with PI3Kδ inhibitors. Paired two-sided Student's *t*-test, Nalm-6: $n = 3$ biological replicates, $P = 0.0088$; RCH-ACV: $n = 6$ biological replicates, $P = 0.0026$. **f**, Western blot of MLC2 levels in Nalm-6 after treatment with α6 integrin-neutralizing antibodies. $n = 3$ independent experiments. p-MLC2, phosphorylated-myosin light chain 2 (Thr18/Ser19). For gel source data, see Supplementary Fig. 1. Scale bars, 100 μm.

To examine whether α6 integrin is a viable candidate regulator of invasion of the CNS by ALL cells, we first assayed its expression in our xenograft models. Nalm-6, RCH-ACV and primary human ALL cells are strongly positive for membrane expression of α6 integrin in vitro and in vivo (Fig. 2d and Extended Data Figs. 4d and 5a, b). We next confirmed that PI3Kδ inhibition with GS-649443 decreases the cell surface expression of α6 integrin by ALL cells (Fig. 2e). Moreover, treatment of Nalm-6 with α6 integrin-neutralizing antibodies decreased phosphorylated MLC2 levels, suggesting that α6 integrin signalling could directly modulate cytoskeletal responses in ALL cells (Fig. 2f). Taken together, these data suggest a feedforward pathway by which PI3Kδ and α6 integrin control migration of ALL cells along laminin through regulation of actomyosin contractility.

ALL cells fail to breach blood–brain barrier

In the adult nervous system, laminin is localized in the ECM of parenchymal microvessels, meninges, the choroid plexus and peripheral nerve sheaths[20]. Because ALL cells can circulate in the vasculature in high numbers, we hypothesized that their metastasis to the CNS was haematogenous, mediated through the interaction of α6 integrin^{+} ALL cells with laminin^{+} microvessels in the brain. To test this, we engrafted fluorescently labelled ALL cells intravenously in mice and used confocal microscopy to image whole mounts of the brain tissue at serial time points after engraftment. The location of ALL cells with respect to the vasculature was determined by injecting a fluorescent vascular blood-pool agent immediately before imaging. We found that ALL cells quickly arrest at branch points in small microvessels, similar to the CNS metastatic process described in melanoma and lung cancer models[21] that show brain metastasis (Fig. 3a–c and Extended Data Fig. 6a–c). In contrast to solid tumour cells, however, ALL cells fail to enter the brain parenchyma and to form proliferative lesions[22].

To determine whether ALL cells cross the blood–brain barrier through leptomeningeal rather than cerebral parenchymal vasculature, we engrafted mice with fluorescent Nalm-6 cells and performed real-time in vivo microscopy of the leptomeningeal vasculature through thinned skull windows (Extended Data Fig. 7a–d, f and Supplementary Videos 1–5). Although ALL cells circulated through and transiently arrested inside leptomeningeal vasculature on the day of intravenous engraftment, Nalm-6 cells did not breach the leptomeningeal blood–brain barrier during this or subsequent post-engraftment time points (Extended Data Fig. 7b–d, f). By contrast, Nalm-6 cells rapidly invaded via diapedesis through the bone marrow vasculature soon after intravenous injection (Extended Data Fig. 7e). Histochemical staining of brain tissue sections confirmed that Nalm-6 cells were absent from leptomeningeal tissue until late disease stages (Fig. 3d–f and Extended Data Fig. 6d, e).

The choroid plexus vasculature has been reported as a site of entry for normal immune cells during CNS trafficking and for haematogenous metastases of solid tumour cells to the meninges[23,24]. We therefore analysed histologic sections of the choroid plexus for the presence of Nalm-6 cells at multiple time points after engraftment. Nalm-6 cells were very rarely detected within either choroid vessels or choroid tissue at time points ranging from the day of intravenous engraftment to end stage disease (Fig. 3d–f).

ALL cells enter the CNS along emissary vessels

We then turned our attention to identifying an alternative route of invasion. Investigators interested in defining the earliest anatomic location of ALL disease in the CNS have studied histologic brain sections of patients and found the first identifiable ALL cells to be localized within the basement membrane of superficial arachnoid veins[25]. Notably, arachnoid veins have been described to merge directly with emissary veins that pass between the meninges en route to the bone marrow via the adjacent skull or vertebral bodies[26] (Fig. 4a). We thus hypothesized that, rather than metastasizing within the circulation, ALL cells migrated directly from the bone marrow to the CNS surrounding the circulation, along the laminin^{+} ECM of emissary bridging vessels. We also hypothesized that directional migration of ALL cells from the bone marrow through these bony channels was due to a potent chemoattractive stimulus within the CSF. Multiple chemokines are indeed present in the CSF, and a wide variety of their cognate receptors are expressed by ALL cells[27]. With the exception of the Notch-driven CCR7^{+} subset of T-ALL, however, no specific chemokine receptor appears to be neces-

© 2018 Springer Nature Limited. All rights reserved.

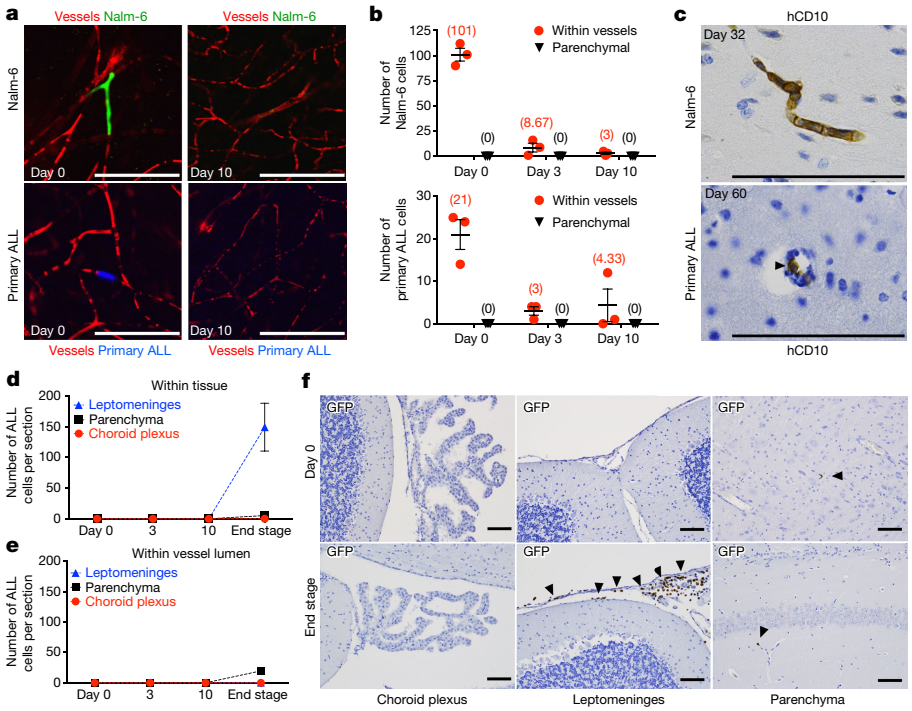


Fig. 3 | ALL cells fail to breach the blood–brain barrier.
a, **b**, Representative confocal microscopy images and quantification of ALL cells located within brain microvessels or parenchyma at various time points after intravenous engraftment. Green, GFP$^+$ Nalm-6 cells; red, vessels indicated by dextran–Alexa Fluor 647; blue, DiR$^+$ primary ALL cells. Data are mean of sample means (cells per mouse) ± s.e.m.; $n = 3$ mice per cell line and time point, 12 micrographs per mouse. **c**, Human (h) CD10$^+$ ALL cells (arrowhead) detected inside lumen of brain parenchymal vessels by immunohistochemistry. Nalm-6: $n = 4$ mice; primary ALL: $n = 10$ mice. **d**, **e**, Quantification of Nalm-6–GFP$^+$ ALL cells within tissue or inside the lumen of blood vessels in the choroid plexus, leptomeninges or brain parenchyma at various time points after intravenous engraftment. Data are mean of sample means (cells per section per mouse) ± s.e.m.; day 0, $n = 3$ mice, 19 sections in total; day 3, $n = 3$ mice, 20 sections; day 10, $n = 3$ mice, 17 sections; end stage disease, $n = 5$ mice, 46 sections. **f**, Representative GFP immunohistochemistry staining of choroid plexus, leptomeninges and brain parenchyma from Nalm-6–GFP-engrafted mice on day 0 after intravenous engraftment ($n = 3$ mice) or at end stage disease ($n = 5$ mice). Arrowheads indicate Nalm-6 ALL cells within leptomeninges or inside the lumen of brain parenchymal microvessels. Scale bars, 100 μm.

sary or sufficient to drive CNS invasion, suggesting that CSF chemokines function additively, synergistically or substitutively to attract ALL cells[27,28]. Consistent with our model, we found that whereas CXCL12 was one of the most highly concentrated chemokines in the CSF (mean, 646 pg ml^{-1}), in vitro invasion of ALL cells towards human CSF is only partially inhibited by AMD3100 blockade of the CXCL12 receptor, CXCR4[29] (Fig. 4b). Moreover, it has previously been shown that in vivo blockade of CXCR4 by continuous infusion AMD3100 does not decrease development of CNS disease in ALL xenografts[27].

We next sought to identify emissary bridging vessels in mice and determine whether ALL cells could be found in transit surrounding these, and whether these vessels were indeed laminin$^+$. As shown in histologic cross-sections in Fig. 4c and Supplementary Videos 6–11, we detected small vessels transiting between the bone marrow and subarachnoid space in control mice. In these locations in leukaemic mice, we found numerous openings in the vertebral cortical bone that were filled with ALL cells, which appeared to be in transit between the involved bone marrow and subarachnoid space (Fig. 4c, d and Supplementary Videos 8–11). Immunohistochemical staining confirmed that these channels contained αSMA$^+$ (vascular smooth muscle) and laminin$^+$ vasculature (Fig. 4e).

Laminin substantially increases in vitro migration of NSPCs in comparison to other ECM substrates[15]. We found that laminin similarly enhanced ALL motility in in vitro invasion assays (Extended Data Fig. 8a, b). Moreover, ALL migration to human CSF along laminin was α6 integrin-dependent and blocked by MLCK inhibition or GS-649443 treatment (Fig. 5a and Extended Data Fig. 8b). Providing further support that ALL migration utilizes α6 integrin–laminin-mediated mechanisms, we found that the frequency of ALL infiltrates along the bone marrow–CNS vascular corridors was significantly decreased in mice treated with GS-649443 (Fig. 5b).

α6 integrin enables abluminal invasion

To specifically examine the role of α6 integrin in CNS invasion of ALL cells in vivo, Nalm-6-engrafted mice were weekly treated with α6 integrin-neutralizing antibodies beginning on day 1 after engraftment until attainment of a clinical end point (Extended Data Fig. 9a). Although there was no difference in peripheral disease burden between targeted and isotype control antibody-treated mice, anti-α6 integrin-treated mice showed a modest increase in survival, and none developed lower extremity paresis by the time of euthanasia (Extended Data Fig. 9b–d). By contrast, 100% of isotype control-treated mice required euthanasia due to paralysis (Fig. 5c). Consistent with this observation, CSF blast counts in anti-α6 integrin-treated mice were reduced by approximately one third (Fig. 5d).

To further address the role of α6 integrin in CNS invasion, we studied the relationship between α6 integrin expression on ALL cells and the incidence of CNS disease in xenograft models and in a retrospective, case–control cohort of patients with ALL who had CNS relapse. Mice engrafted with ALL cells that express high surface levels of α6 integrin (>25% cells positive by flow cytometry and 2–3+ staining intensity by immunohistochemistry) displayed a higher frequency of CNS disease at clinical end point compared to mice engrafted with ALL cells that expressed low levels of α6 integrin (<10% cells positive by flow cytometry and 0–1+ staining intensity by immunohistochemistry; Fig. 5e and Extended Data Fig. 5b, c). In archived bone marrow biopsies from patients with ALL who either did or did not develop CNS metastases, α6 integrin expression on ALL cells was associated with the occurrence of CNS relapse (Fig. 5f, g). Expression of α6 integrin in this patient cohort was independent of other variables considered to increase the risk of CNS disease involvement (Extended Data Table 1).

© 2018 Springer Nature Limited. All rights reserved.

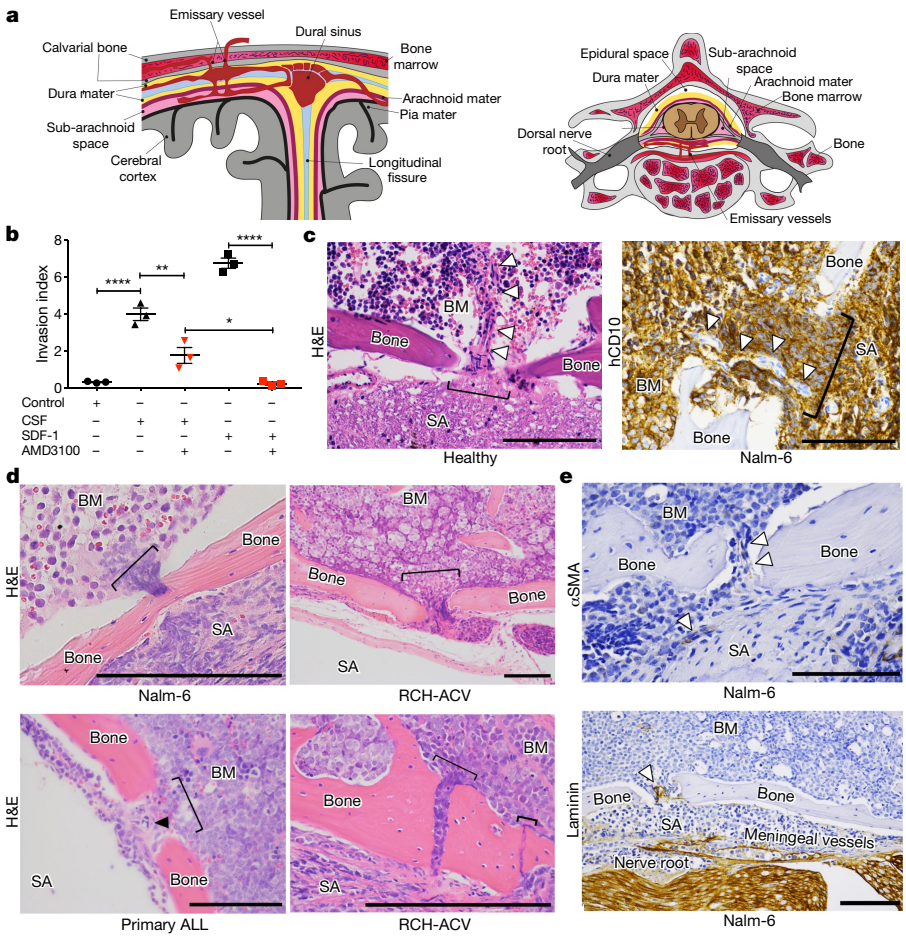


Fig. 4 | ALL cells passage to the CNS subarachnoid space along the abluminal surface of laminin$^+$ bone marrow emissary vessels. a, Anatomy of emissary vessels in calvarial and vertebral bone marrow. **b**, ALL in vitro invasion towards human CSF or CXCL12 (also known as SDF1) with or without CXCR4 blockade (AMD3100). Data are mean ± s.e.m.; ANOVA with Tukey test, $n = 3$ biological replicates, $^*P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$. **c**, Bone channel (bracket) containing emissary vessels (arrowheads) connecting the vertebral bone marrow (BM) with subarachnoid space (SA) in histologic sections of vertebrae from healthy ($n = 4$ mice) and Nalm-6-engrafted (human CD10$^+$ cells) leukaemic mice ($n = 3$ mice). **d**, ALL in transit to the CNS through bone channels. Vertebrae of mice engrafted with primary ALL, Nalm-6 and RCH-ACV cells are shown. Brackets, ALL$^+$ bone channels; arrowhead, emissary vessels. Nalm-6: $n = 27$ mice, primary ALL: $n = 13$ mice, RCH-ACV: $n = 10$ mice. **e**, αSMA (vascular smooth muscle, $n = 26$ mice) and laminin ($n = 7$ mice) immunohistochemistry staining of emissary vessels (indicated by the arrowheads). Scale bars, 100 μm.

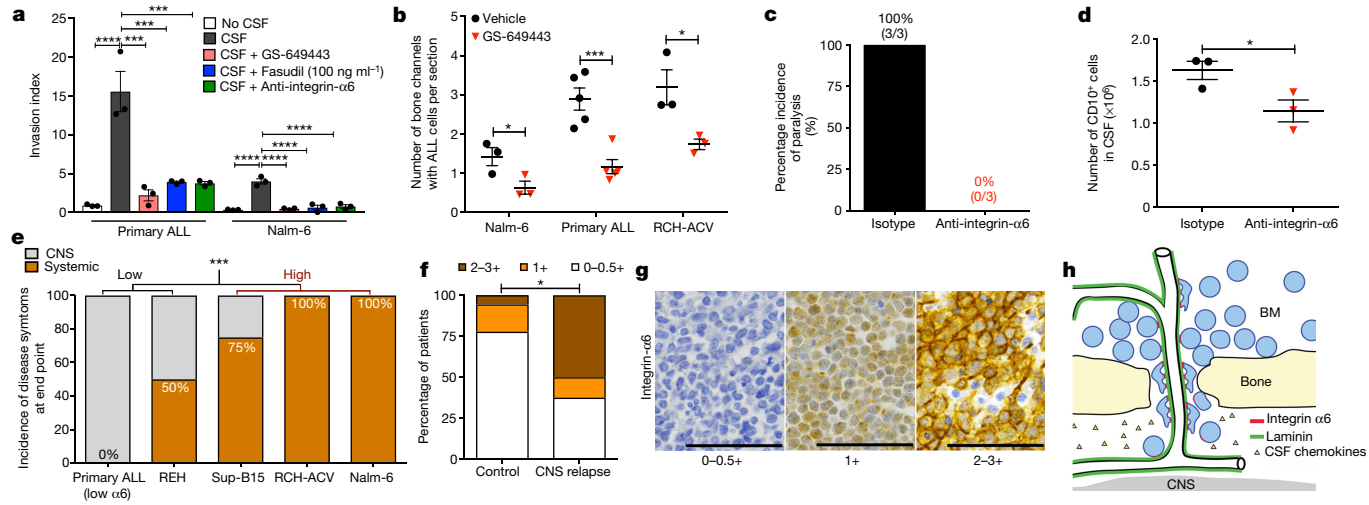


Fig. 5 | ALL cells use α6 integrin–laminin-dependent interactions to invade the CNS. a, ALL invasion towards human CSF along laminin–collagen matrices after treatment with α6 integrin-neutralizing antibodies, the Rho-kinase inhibitor fasudil (at MLCK inhibitory dose), or GS-649443. Data are mean ± s.e.m.; ANOVA with Tukey, $n = 3$ biological replicates, $^{***}P < 0.001$, $^{****}P < 0.0001$. **b**, Frequency of ALL cell infiltration in bony channels of the spine in vehicle- versus GS-649443-treated mice. Data are mean ± s.e.m.; unpaired two-sided Student's *t*-test, Nalm-6: $n = 3$ vehicle-treated mice, 4 GS-649443-treated mice, minimum of 24 vertebral sections per mouse, $P = 0.0198$; primary human ALL: $n = 5$ mice per treatment group, minimum of 10 vertebral sections per mouse, $P = 0.0009$; RCH-ACV: $n = 4$ mice per treatment group, minimum of 8 vertebral sections per mouse, $P = 0.0149$. **c**, Incidence of hindlimb paralysis at time of euthanasia. **d**, CNS disease burden (number of CD10$^+$ ALL cells) in isotype control and α6 integrin-neutralizing antibody-treated Nalm-6-engrafted mice at euthanasia for matched time points. Data are mean ± s.e.m.; unpaired two-sided Student's *t*-test, $n = 3$ mice per treatment group, $P = 0.0234$. **e**, Incidence of CNS disease symptoms (hindlimb paralysis) at clinical end point in mice engrafted with ALL cells expressing low versus high levels of integrin α6. Fisher's exact test, primary human ALL integrin α6 low, REH and RCH-ACV: $n = 6$ mice each; SUP-B15: $n = 3$ mice; Nalm-6: $n = 10$ mice, $P = 0.0008$. **f**, Association between α6 integrin expression in bone marrow blasts and CNS disease relapse in patients with ALL. Freeman–Halton extension of Fisher's exact test, no CNS relapse, $n = 18$ patients, CNS relapse, $n = 8$ patients, $P = 0.0282$. **g**, Representative α6 integrin immunohistochemistry of bone marrow biopsy samples from patients (0–0.5+: $n = 17$ patients; 1+: $n = 4$ patients; 2+–3+: $n = 5$ patients. **H**, Schematic of ALL CNS invasion model. Scale bars, 50 μm.

© 2018 Springer Nature Limited. All rights reserved.

There are currently no molecularly targeted interventions to prevent or treat CNS disease in ALL, reflecting a limited understanding of the mechanisms of ALL metastasis to the CNS. Here we show that ALL cells use a NSPC embryonic pathfinding mechanism in order to traverse the vascular channels that connect the bone marrow and meninges, the predominant site of CNS ALL disease in humans[14,30]. Moreover, we show that this process can be intercepted in mouse xenograft models, providing a rationale for using clinically available PI3Kδ inhibitors to prevent CNS disease involvement in patients.

Leukaemic cells have previously been shown to hijack haematopoietic stem cell trafficking mechanisms to metastasize within the bone marrow microenvironment[10]. Our data demonstrate that this pattern of molecular plagiarism by ALL cells extends across organ systems, potentially to exploit a path of least resistance into the CNS. By travelling along the external (abluminal) surface of vessels that are topologically contiguous with the CNS subarachnoid space, ALL cells migrate directly from the bone marrow to the CNS, bypassing the need to enter and exit the CNS vasculature (Fig. 5h). In so doing, ALL cells phenocopy α6 integrin-dependent mechanisms used by NSPCs to migrate to the olfactory bulb along the ECM of vessels[31]. In future studies, it will be important to examine whether ALL cells similarly utilize other integrin–ECM interactions that drive neuronal pathfinding and developmental organization in order to the metastasize to the CNS[32].

Although the CNS is widely considered to be an immune-privileged site, the recent discovery in mice of a lymphatic system in the dura mater that drains macromolecules and immune cells from the brain suggests that other unrecognized pathways might permit communication between the periphery and the CNS[33]. Future studies may reveal that this unique ALL trafficking pathway is involved in immune surveillance or inflammatory processes. Exploring the interactions between normal and malignant immune cells and these vascular scaffolds may thus reveal multiple points of intervention to treat CNS invasive processes.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0342-5.

Received: 20 September 2017; Accepted: 12 June 2018;
Published online 18 July 2018.

1. Garcia-Manero, G., Kantarjian, H. M. & Schiffer, C. A. in *Holland-Frei Cancer Medicine* (eds Kufe, D. W. et al.) (BC Decker, Hamilton, 2003).
2. Surapaneni, U. R. et al. Central nervous system relapse in adults with acute lymphoblastic leukemia. *Cancer* **94**, 773–779 (2002).
3. Vanhaesebroeck, B., Guillermet-Guibert, J., Graupera, M. & Bilanges, B. The emerging mechanisms of isoform-specific PI3K signalling. *Nat. Rev. Mol. Cell Biol.* **11**, 329–341 (2010).
4. Lannutti, B. J. et al. CAL-101, a p110δ selective phosphatidylinositol-3-kinase inhibitor for the treatment of B-cell malignancies, inhibits PI3K signaling and cellular viability. *Blood* **117**, 591–594 (2011).
5. Furman, R. R. et al. Idelalisib and rituximab in relapsed chronic lymphocytic leukemia. *N. Engl. J. Med.* **370**, 997–1007 (2014).
6. Hanna, S. & El-Sibai, M. Signaling networks of Rho GTPases in cell motility. *Cell. Signal.* **25**, 1955–1961 (2013).
7. Parri, M. & Chiarugi, P. Rac and Rho GTPases in cancer cell motility control. *Cell Commun. Signal.* **8**, 23 (2010).
8. Reiske, H. R. et al. Requirement of phosphatidylinositol 3-kinase in focal adhesion kinase-promoted cell migration. *J. Biol. Chem.* **274**, 12361–12366 (1999).
9. Webb, D. J. et al. FAK–Src signalling through paxillin, ERK and MLCK regulates adhesion disassembly. *Nat. Cell Biol.* **6**, 154–161 (2004).
10. Sipkins, D. A. et al. In vivo imaging of specialized bone marrow endothelial microdomains for tumour engraftment. *Nature* **435**, 969–973 (2005).
11. Colmone, A. et al. Leukemic cells create bone marrow niches that disrupt the behavior of normal hematopoietic progenitor cells. *Science* **322**, 1861–1865 (2008).
12. Boyerinas, B. et al. Adhesion to osteopontin in the bone marrow niche regulates lymphoblastic leukemia cell dormancy. *Blood* **121**, 4821–4831 (2013).
13. Mercurio, A. M., Rabinovitz, I. & Shaw, L. M. The α6β4 integrin and epithelial cell migration. *Curr. Opin. Cell Biol.* **13**, 541–545 (2001).
14. Georges-Labouesse, E., Mark, M., Messaddeq, N. & Gansmüller, A. Essential role of α6 integrins in cortical and retinal lamination. *Curr. Biol.* **8**, 983–986 (1998).
15. Flanagan, L. A., Rebaza, L. M., Derzic, S., Schwartz, P. H. & Monuki, E. S. Regulation of human neural precursor cells by laminin and integrins. *J. Neurosci. Res.* **83**, 845–856 (2006).
16. Belkin, A. M. & Stepp, M. A. Integrins as receptors for laminins. *Microsc. Res. Tech.* **51**, 280–301 (2000).
17. DiGiuseppe, J. A., Fuller, S. G. & Borowitz, M. J. Overexpression of CD49f in precursor B-cell acute lymphoblastic leukemia: potential usefulness in minimal residual disease detection. *Cytometry B Clin. Cytom.* **76B**, 150–155 (2009).
18. Coustan-Smith, E. et al. New markers for minimal residual disease detection in acute lymphoblastic leukemia. *Blood* **117**, 6267–6276 (2011).
19. Terol, M.-J. et al. Expression of beta-integrin adhesion molecules in non-Hodgkin's lymphoma: correlation with clinical and evolutive features. *J. Clin. Oncol.* **17**, 1869–1875 (1999).
20. Colognato, H., ffrench-Constant, C. & Feltri, M. L. Human diseases reveal novel roles for neural laminins. *Trends Neurosci.* **28**, 480–486 (2005).
21. Kienast, Y. et al. Real-time imaging reveals the single steps of brain metastasis formation. *Nat. Med.* **16**, 116–122 (2010).
22. Takei, H., Rouah, E. & Ishida, Y. Brain metastasis: clinical characteristics, pathological findings and molecular subtyping for therapeutic implications. *Brain Tumor Pathol.* **33**, 1–12 (2016).
23. Wilson, E. H., Weninger, W. & Hunter, C. A. Trafficking of immune cells in the central nervous system. *J. Clin. Invest.* **120**, 1368–1379 (2010).
24. Boire, A. et al. Complement component 3 adapts the cerebrospinal fluid for leptomeningeal metastasis. *Cell* **168**, 1101–1113 (2017).
25. Price, R. A. & Johnson, W. W. The central nervous system in childhood leukemia. I. The arachnoid. *Cancer* **31**, 520–533 (1973).
26. Leestma, J. E. *Forensic Neuropathology* 2nd edn (CRC, Boca Raton, 2008).
27. Williams, M. T. et al. The ability to cross the blood–cerebrospinal fluid barrier is a generic property of acute lymphoblastic leukemia blasts. *Blood* **127**, 1998–2006 (2016).
28. Buonamici, S. et al. CCR7 signalling as an essential regulator of CNS infiltration in T-cell leukaemia. *Nature* **459**, 1000–1004 (2009).
29. Kothur, K., Wienholt, L., Brilot, F. & Dale, R. C. CSF cytokines/chemokines as biomarkers in neuroinflammatory CNS disorders: a systematic review. *Cytokine* **77**, 227–237 (2016).
30. Radmanesh, F. et al. Mutations in LAMB1 cause cobblestone brain malformation without muscular or ocular abnormalities. *Am. J. Hum. Genet.* **92**, 468–474 (2013).
31. Bovetti, S. et al. Blood vessels form a scaffold for neuroblast migration in the adult olfactory bulb. *J. Neurosci.* **27**, 5976–5980 (2007).
32. Wade, A., McKinney, A. & Phillips, J. J. Matrix regulators in neural stem cell functions. *Biochim. Biophys. Acta* **1840**, 2520–2525 (2014).
33. Louveau, A. et al. Structural and functional features of central nervous system lymphatic vessels. *Nature* **523**, 337–341 (2015).

Acknowledgements We thank the Duke Sequencing and Genomic Technologies (Duke Cancer Institute) and Genomic and Computational Biology shared resources and H. Dressman for gene array processing and analysis services and technical advice; M. K. Nicholas for providing expert opinions on neuro-oncology, Z. Li for assistance with statistical analysis and C. Queva for helpful discussions regarding PI3K inhibition. This work was supported by the Duke Cancer Institute and Gilead Sciences, Inc.

Reviewer information *Nature* thanks I. Aifantis, J. Burger and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.A.S. and H.Y. designed the study and analysed the data. D.A.S. conceived and supervised the project and compiled patient data. H.Y. carried out animal experiments, pharmacokinetic studies, flow cytometry, western blotting and transwell migration assays. T.T.P. performed animal experiments, confocal microscopy and flow cytometry, analysed data and assembled figures. G.C. performed invasion assays, western blotting, flow cytometry and three-dimensional reconstructions, analysed data, assembled figures and assisted with animal experiments. B.N. and L.O. assisted with animal studies and performed histology and immunohistochemistry. M.J.W. performed microarray analyses, bone marrow intravital microscopy and histologic analyses. S.M.R. analysed data, assembled figures, performed transwell migration assays and assisted with animal experiments. E.M.J. performed intravital microscopy through thinned skull windows. S.T. and J.T. provided PI3Kδ inhibitors and assisted with pharmacokinetic analyses. C.M.M. reviewed all histology and scored immunohistochemistry staining of bone marrow biopsy samples obtained from patients. A.C. analysed data and contributed to the figures. D.A.S. wrote the manuscript. H.Y., T.T.P., G.C., S.M.R., S.T., C.M.M. and A.C. edited the paper. Key findings in the paper were replicated by a second postdoctoral researcher in the laboratory.

Competing interests D.A.S. received a grant award from Gilead Sciences, Inc. S.T. and J.T. are employees of Gilead Sciences, Inc.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0342-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0342-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.A.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized.

Collection of bone marrow, spleen and CSF cells. Mouse femurs were collected and bone marrow cells were aspirated with RPMI 1640 medium containing 10% fetal bovine serum (FBS). Mouse spleen tissues were homogenized in RPMI 1640 medium containing 10% FBS using a 15-ml syringe. The spine was removed from mice and carefully separated into individual vertebral bodies by cutting through the intervertebral discs. The spinal cord and meninges of each vertebral body were then washed with RPMI 1640 containing 10% FBS to collect cells from the CSF. The bone marrow, spleen and CSF cells were passed through 70-μm filters, washed with phosphate-buffered saline (PBS) and treated with ACK lysis buffer to remove red blood cells.

Flow cytometry analysis. For in vitro cell surface receptor staining, 1×10^6 cells per 100 μl were stained with antibody for 30 min at 4 °C in Automacs buffer (BD Biosciences) containing 3% BSA, washed with 3% BSA, and analysed on a FACS Canto II cytometer. For apoptosis analysis, cells were resuspended in Annexin V Binding Buffer (BD Biosciences) and stained with anti-annexin V–AF647 antibody and propidium iodide for 15 min at room temperature and analysed. For cell cycle analysis, cells were first stained for CD10 and then fixed in 2% paraformaldehyde for 30 min at 4 °C, permeabilized in 80% ethanol overnight at −20 °C, washed and resuspended in Automacs buffer containing 3% BSA and then incubated with anti-human-KI-67 for 1 h at 4 °C before washing with 3% BSA before analysis. In some experiments, cell cycle analysis was performed by staining 1×10^6 cells per 100 μl Automacs buffer containing 1% FBS with 4 μg ml^{-1} Hoechst 34580 (Invitrogen) for 30 min at room temperature. An example of our gating strategy is shown in Supplementary Fig. 2.

Labelling of cells using fluorescent dyes. Cells were fluorescently labelled through incubation with 'DiR' lipophilic dyes (DiIC1 8(7) (1,1′-dioctadecyl-3,3,3′,3′-tetramethylindotricarbocyanine iodide, Invitrogen) as previously described[10]. In brief, cells were isolated at a density of 2.5×10^6 cells per ml and resuspended in 25 μM DiR in culture medium containing 10% FBS. Cells were protected from the light and rotated at room temperature for 30 min. After incubation, cells were washed twice in PBS, counted and resuspended at desired concentrations for engraftment. Dye labelling efficiency was tested using flow cytometry.

Mouse engraftment. Specific pathogen-free 6- to 8-week-old male and female SCID mice (Charles River) were inoculated intravenously with 5×10^6 Nalm-6–GFP, RCH-ACV or REH cells in PBS through the tail vein. Similarly, primary ALL or SUP-B15 cells were engrafted into 6- to 8-week-old male and female NOD/SCID/IL-2Rγ (NSG) mice (Charles River) by intravenous injection. All experimental procedures involving mice were approved by the Animal Care and Use Committee of Duke University. The Institutional Review Board of Duke University approved the use of de-identified primary human ALL cells in this study (Protocol 00006268, 19-April-08). All experiments were performed in accordance with the relevant ethical guidelines and regulations.

Administration of GS-649443. GS-649443 or vehicle (Gilead Sciences, Inc.) was administrated to Nalm-6–GFP, RCH-ACV, REH, SUP-B15 or primary ALL cell-engrafted mice by oral gavage from day 1 or day 20 after engraftment. Mice were dosed twice a day with 2 mg kg−1 GS-649443. Mice were monitored daily and euthanized when they showed hind limb paralysis or other observable CNS symptoms, severe cachexia (≥20% weight loss), respiratory or other distress, or extreme lethargy. For paired-animal analysis, mice from the vehicle- and GS-649443-treated groups were paired at the beginning of the study. Both mice in each pair were euthanized when either of them showed any of the symptoms outlined above.

Administration of α6 integrin-neutralizing antibodies. Anti-human α6 integrin-neutralizing antibodies (Invitrogen) or isotype control antibodies (rat gamma globulin) (Jackson ImmunoResearch) were administrated intravenously in Nalm-6–GFP cell-engrafted mice. The mice were dosed with antibodies weekly at 3 mg kg^{-1}. Mice were monitored daily and euthanized at the clinical end points described above.

Cell lines and culture. The Nalm-6 (ACC-128) and RCH-ACV (ACC-548) cell lines were purchased from DSMZ. REH (CRL-8286) and SUP-B15 (CRL-1929) cell lines were purchased from ATCC. STR profiling was repeated routinely to authenticate cell lines. Mycoplasma testing by PCR was performed every three months on all cell lines in culture. All mycoplasma testing by PCR was negative during the entire duration of this study. The GFP-expressing clone of Nalm-6 was generated as described previously[11]. Nalm-6, REH and RCH-ACV cells were cultured in RPMI 1640 (Corning Inc.) supplemented with 10% FBS (Gemini Bio-Products). Sup-B15 cells were cultured in Iscove's modified Dulbecco's medium (IMDM), supplemented with 20% FBS and 0.05 mM 2-mercaptoethanol. All cultures were maintained at 37 °C in a 5% CO_2 humidified atmosphere.

Antibodies for flow cytometry. The following antibodies were purchased from BD Pharmingen: an allophycocyanin (APC)-conjugated mouse antibody against human CD10, peridium chlorophyll protein complex with cyanin-5.5 (PerCP–Cy5.5)-conjugated mouse antibody against human CD10, phycoerythrin (PE)-conjugated rat antibodies against human CXCR4, APC-conjugated mouse antibody against human CXCR3, APC-conjugated mouse antibody against human CD44, PE-conjugated mouse antibodies against human CD24, APC-conjugated rat antibody against mouse CD45, PE-conjugated rat antibody against human CD49f, Alexa Fluor647-conjugated annexin-V antibody, PE-conjugated KI-67 antibody and PE-conjugated mouse antibody against human CD19.

Measurement of GS-649443 concentration in plasma and CNS. Brain tissues were collected after perfusion of mice with saline. The thoracic cavity of the mouse was opened and saline was perfused at a rate of 5 ml min^{-1} with a 23G needle through the left cardiac ventricle for 6 min. Collected brain tissues were homogenized, diluted three times and treated with acetonitrile. Plasma samples were diluted twice and treated with acetonitrile. The above mixtures were centrifuged at 4,275 r.p.m. for 15 min and their supernatants were transferred to a 96-well plate for analysis using the API 5000 triple quadrupole mass spectrometer (AB Sciex).

Transwell migration assay. Migration assays were performed using 24-well plates containing 0.3% FBS in medium. Nalm-6–GFP cells were treated with vehicle, GS-649443 (100 nM), AZD5363 (10 μM) or AMD3100 (5 μg ml^{-1}) for 20 min in serum-free medium and seeded on an uncoated polycarbonate membrane insert (6.5 mm in diameter with 8.0-μm pores) in a Transwell apparatus (Corning). The lower chamber of the Transwell plate was loaded with 600 μl of medium with 0.3% serum alone or CXCL12 (100 ng ml^{-1}) as a positive control. After incubation for 3 h at 37 °C, the inserts were removed and cells that had migrated to the bottom were counted using a haemocytometer.

Western blot analysis. Following treatment with GS-649443 or anti-α6 integrin, more than 1×10^6 cells were collected by centrifugation and washed with ice-cold PBS. Ice-cold RIPA buffer (50 mM Tris-HCl (pH 7.4), 0.25 M NaCl, 5 mM EDTA, 20 mM NaF, 1% NP-40) containing fresh protease inhibitor cocktail (Sigma-Aldrich) and phosphatase inhibitor cocktails 2 and 3 (Sigma-Aldrich) was added to the cells. The suspension was transferred into a centrifuge tube and placed on ice for 3 min. The cell suspension was cleared by centrifugation at 14,000*g* for 3 min at 4 °C. The supernatants (total cell lysate) were used immediately or stored at −80 °C. Protein concentrations were determined using the DC Protein Assay (Bio-Rad Laboratories). Samples (20 μg of protein) were analysed using the following primary antibodies, as indicated: anti-phosphorylated-myosin light chain 2 (Thr18/Ser19) (Cell Signaling Technology), anti-myosin light chain 2 (Cell Signaling Technology), anti-β-actin (Abcam). Horseradish peroxidase (HRP)-coupled rabbit IgG (Cell Signaling Technology) and goat anti-mouse IgG–HRP (ThermoFisher) were used as secondary antibodies, and immunoreactive proteins were detected by enhanced chemiluminescence (ECL) (ThermoFisher).

Three-dimensional invasion assays. Cells in 0.3% serum were treated with GS-649443 (100 nM) for 16 h, or cultured in 0.3% serum overnight before the addition of anti-human α6 integrin-neutralizing antibodies for 3 h (Invitrogen, 40 μg ml^{-1}) or AMD3100 (5 μg ml^{-1}) for 20 min. Cells were resuspended in serum-free rat-tail collagen (Advanced Biomatrix 5153 at 3 mg ml^{-1}) alone or supplemented with laminin (Sigma-Aldrich, L2020) or fibronectin (Sigma-Aldrich, F1141). Wherever unspecified, cells were resuspended in collagen and laminin (0.01 mg ml^{-1}). Resuspended cells were aliquoted into 96-well plates and spun down to the bottom of the plate. Collagen was allowed to polymerize for 2 h and cell culture medium or human CSF (LEE Biosolutions, 991-19-P) was added on top of the gel as a chemoattractant. After 16 h of incubation at 37 °C, plates were fixed and stained with 5 μg ml^{-1} Hoechst 33258 (Molecular Probes–Life Technologies). Plates were imaged on a Zeiss 780 inverted confocal microscope (Carl Zeiss, Germany) with Zen software. The three-dimensional migration index was calculated as number of invading cells at 50 μm divided by the total number of cells.

Immunohistochemistry. CD10, αSMA, laminin, α6 integrin and GFP immunohistochemistry staining were performed under the following conditions. CD10: antigen retrieval was performed using EDTA Target Retrieval solution (DAKO). Slides were stained with anti-CD10 (clone EPR5904-110; Abcam) at 1:4,000. DAKO Envision System-HRP (DAB), for use with rabbit (K4011; DAKO), was used as a secondary antibody. αSMA: no antigen retrieval was performed. Slides were stained with anti-β-actin, α-smooth muscle actin (clone 1A4; Sigma-Aldrich) at a concentration of 3 μg ml^{-1}. M.O.M. Biotinylated Anti-Mouse IgG Reagent, for use with mouse (PK-2200; Vector Laboratories), was used as a secondary antibody. Laminin: antigen retrieval was performed using Citrate Target Retrieval solution (DAKO). Slides were stained with anti-laminin (clone ab11575; Abcam) at a concentration of 5 μg ml^{-1}. DAKO Envision System-HRP (DAB), for use with rabbit (K4011; DAKO), was used as a secondary antibody. α6 integrin: Antigen retrieval was performed using EDTA Target Retrieval solution (DAKO). Slides were stained with anti-integrin α6 (clone EPR18124; Abcam) at a concentration of 6 μg ml^{-1}. DAKO Envision System-HRP (DAB), for use with rabbit (K4011; DAKO), was used as a secondary antibody. GFP: slides were stained with anti-GFP (NB600-308; NOVUS Biologicals) at a concentration of 1 g ml^{-1}. DAKO Envision System-HRP (DAB), for use with rabbit (K4011; DAKO), was used as a secondary

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

antibody. Nalm-6–GFP$^+$ ALL cell counts in individual GFP-stained brain sections were performed by a trained haematopathologist (C.M.M.). Scoring of ALL cell α6 integrin expression in mouse bone marrow specimens was performed in a blinded fashion by a trained hematopathologist (C.M.M.).

Three-dimensional reconstructions. Three-dimensional reconstructions of emissary vessels were acquired from 30-μm sections of the spinal cord stained with eosin and haematoxylin. Consecutive images were acquired at regular intervals and assembled into movies using VirtualDub64 software.

Quantification of vertebral bone channels containing ALL cells. First, 5-μm sections of the spinal cord were stained with eosin and haematoxylin. Then, vertebral bone channels that contained ALL cells were identified and counted in each intact spinal section. This analysis was performed in a blinded fashion by two independent laboratory members.

Human samples. The study design and use of clinical bone marrow biopsy samples was approved by the Institutional Review Board of Duke University (protocol 57532). A waiver of informed consent was granted as the research was considered to present no more than minimal risk of harm to participants. Electronic medical records of the Duke University Health System from 2009 to 2016 were searched using the Duke Enterprise Data Unified Content Explorer (DEDUCE) online query system in order to identify patients of any age with ALL diagnoses. The medical chart of each patient was then reviewed to confirm the pathologic diagnosis as well as the availability of archived diagnostic and/or relapse bone marrow biopsy specimens. Patients with ALL with CNS relapse confirmed through identification of malignant cells in the CSF by cytology and/or immunophenotyping were retrospectively enrolled. Patients with ALL who did not undergo bone marrow transplantation and who were without relapse in the bone marrow or CNS after more than 5 years of follow-up served as control patients. Patients with ALL who did not undergo bone marrow transplantation and who relapsed in the bone marrow but had no clinical evidence of CNS relapse and confirmed negative CSF cytology/immunophenotyping served as additional control patients. Scoring of α6 integrin expression in ALL cells in imunohistochemically stained bone marrow biopsies was performed by a trained haematopathologist.

Intravital microscopy. SCID mice were anaesthetized using isoflurane and a rectangular incision was made in the scalp, revealing the intact, underlying cortical bone. The region was washed with PBS and fluorescently labelled dextran (Dex–FITC or Dex–AF647) was administered via tail vein injection to highlight the vasculature. Dex-AF647 is a large molecular weight dextran molecule conjugated to the fluorophore Alexa Fluor647. Because it has a prolonged blood-pool half-life, it is a useful means of visualizing the vasculature in in vivo studies. Mice were placed in a specially designed restrictor and a cover slip was placed over the exposed calvarial bone. Mice remained anaesthetized throughout the procedure. High-resolution images were obtained through the intact mouse skull using a Leica SP5 confocal and multiphoton microscope with a 20×/0.40 numerical aperture (NA) objective lens. The system utilizes a femtosecond titanium:sapphire laser (Chameleon) for multi-photon or single-photon excitation and multiple Cs lasers (including an Argon laser, a HeNe laser and 561- and 633-nm diode lasers) for single-photon excitation. Images were captured using Leica LAS-AF software using line and frame averaging. The calvarial bone marrow was subdivided into numbered anatomical areas and overlapping 20× images of the entire region were captured. After the procedure, these images were merged using Photoshop v.12.0.4 to generate a montage image of the entire calvarium. Cells were then enumerated using the cell counter function in ImageJ.

Skull thinning. SCID mice were anaesthetized using isoflurane and the cortical bone was exposed following a sagittal incision in the scalp. The head of the mouse was secured using a stereotactic apparatus. A drill (Hager and Meisinger GMBH size 1007, through Fine Surgical Tools) was used to thin the skull, creating a small, circular bowl-like area with a flat bottom. A microsurgical blade was used to carefully complete the thinning by hand (Surgistar, 38–6961). High-resolution images of the leptomeninges were obtained through the thinned skull window using a Leica SP5 confocal and multi-photon microscope with a 20×/0.40 NA objective lens.

Confocal microscopy brain whole-mount imaging. SCID mice were engrafted with 5×10^6 DiR-labelled cells and confocal imaging of the brain was performed at indicated time points after engraftment (day 0, 2 h after engraftment; day 3; day 10). Mice were euthanized at imaging time points, the brain was removed and then cut sagittally along the midline into thin sections for whole-mount imaging. High-resolution images of ALL cells in brain parenchymal spaces was obtained using a Leica SP5 confocal and multi-photon microscope with a 20 ×/0.40 NA objective lens. The system uses a femtosecond titanium:sapphire laser (Chameleon) for multi-photon or single-photon excitation and multiple Cs lasers (including an argon laser, a HeNe laser, and 561- and 633-nm diode lasers) for single-photon excitation. Images were captured with Leica LAS-AF software.

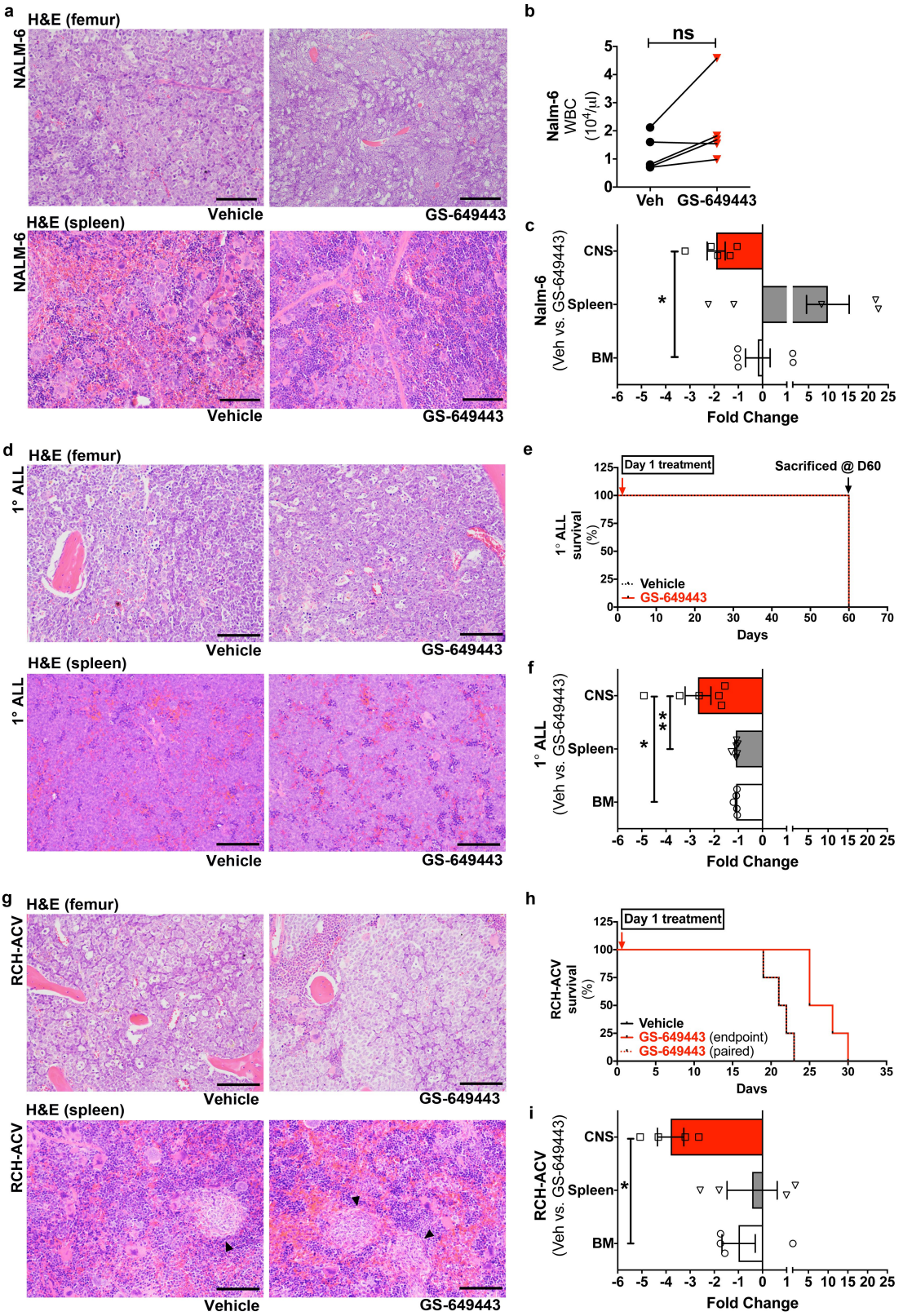
Microarray analysis. Total RNA isolated from bone marrow and CSF was sent for gene expression profiling using the Clariom D Human microarrays (Affymetrix) at the Duke Center for Genomic and Computational Biology. Raw cell intensity data were imported in to the Expression Console (Affymetrix) and bone marrow arrays (vehicle and GS-649443 samples) and CSF arrays (vehicle and GS-649443) were normalized using the RMA algorithm. Gene level differential expression analysis was performed using the Transcript Analysis Console (v.3.1.0.5; Affymetrix) and ANOVA statistical analyses were performed on the bone marrow and CSF arrays separately. Transcripts identified to be differentially regulated in the bone marrow or CSF samples by ±1.5-fold with $P < 0.05$ were carried forward for pathway analysis.

Statistics and reproducibility. Microarray analysis was performed as described above. Kaplan–Meier curves with two-sided log rank Mantel–Cox analysis were used to assess in vivo survival. Two-sided paired Student's *t*-tests were used for analysis of vehicle versus inhibitor effects on tumour burden in mice, ALL bone marrow migration in mice, and GS-649443 effects on α6 integrin expression in ALL cells in vitro. One-way ANOVA analyses with Tukey post hoc testing for multiple comparisons were performed for invasion assays, transwell migration assays and comparison of tumour burden across bone marrow, spleen and CNS in vehicle versus inhibitor-treated mice. Fisher's exact tests and the Freeman–Halton extension of the Fisher's exact test were used to compare α6 integrin staining levels in human biopsies and on CNS symptoms incidence levels. Nalm-6 migration along calvarial bone marrow vessels was quantified using Image J software (NIH). The majority of statistical analysis was performed using GraphPad Prism version 7.0, while the VassarStats statistical computation website was used to perform the Freeman–Halton extension of the Fisher's exact test. Significant *P* values were defined as follows: NS, not significant; $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$. Precise *P* values for data shown in the main figures using these ranges are as follows. Figure 1e: Nalm-6 xenografts: $P = 0.8142$ (bone marrow), $P = 0.1014$ (spleen). Figure 1g: 1° ALL xenografts: $P = 0.0052$ (CNS), $P = 0.0266$ (bone marrow), 0.0222 (spleen); RCH-ACV xenografts: $P = 0.0067$ (CNS), $P = 0.0981$ (bone marrow), $P = 0.2465$ (spleen). Figure 2b: $P = 0.9657$ (vehicle versus AZD5363), $P = 0.0035$ (vehicle versus idelalisib), $P = 0.0040$ (vehicle versus GS-649443). Figure 2c: $P = 0.1337$ (vehicle versus GS-649443). Figure 4b: $P < 0.0001$ (control versus CSF), $P = 0.0015$ (CSF versus CSF and AMD3100), $P = 0.0201$ (CSF and AMD3100 versus SDF1 and AMD3100), $P < 0.0001$ (SDF1 versus SDF1 and AMD3100); Fig. 5a: primary ALL: $P < 0.0001$ (no CSF versus CSF), $P = 0.0001$ (CSF versus GS-649443), $P = 0.0003$ (CSF versus fasudil), $P = 0.0003$ (CSF versus anti-α6 integrin), Nalm-6: $P < 0.0001$ (no CSF versus CSF), $P < 0.0001$ (CSF versus GS-649443), $P < 0.0001$ (CSF versus fasudil), $P < 0.0001$ (CSF versus anti-α6 integrin). Data are mean ± s.e.m.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Source Data for quantifications mentioned in the text or shown in the graphs plotted in Figs. 1–5 and Extended Data Figs. 1–9 are available in the online version of this paper. Full scans of western blots are provided in Supplementary Fig. 1. Microarray gene expression data that support the findings of this study have been deposited in Gene Expression Omnibus (GEO) under accession number GSE114627. The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information, as Source Data or from the corresponding author upon reasonable request.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

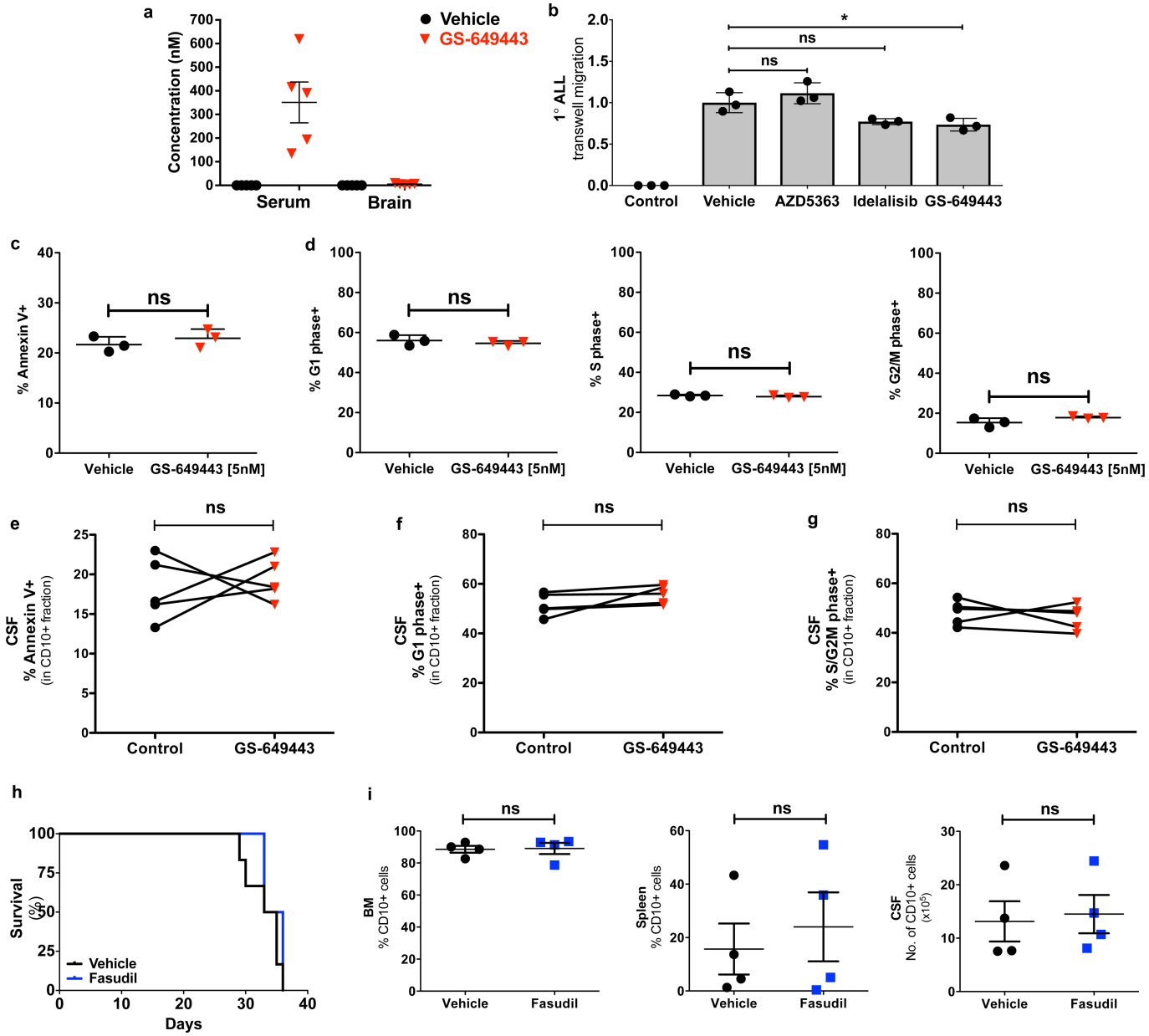


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

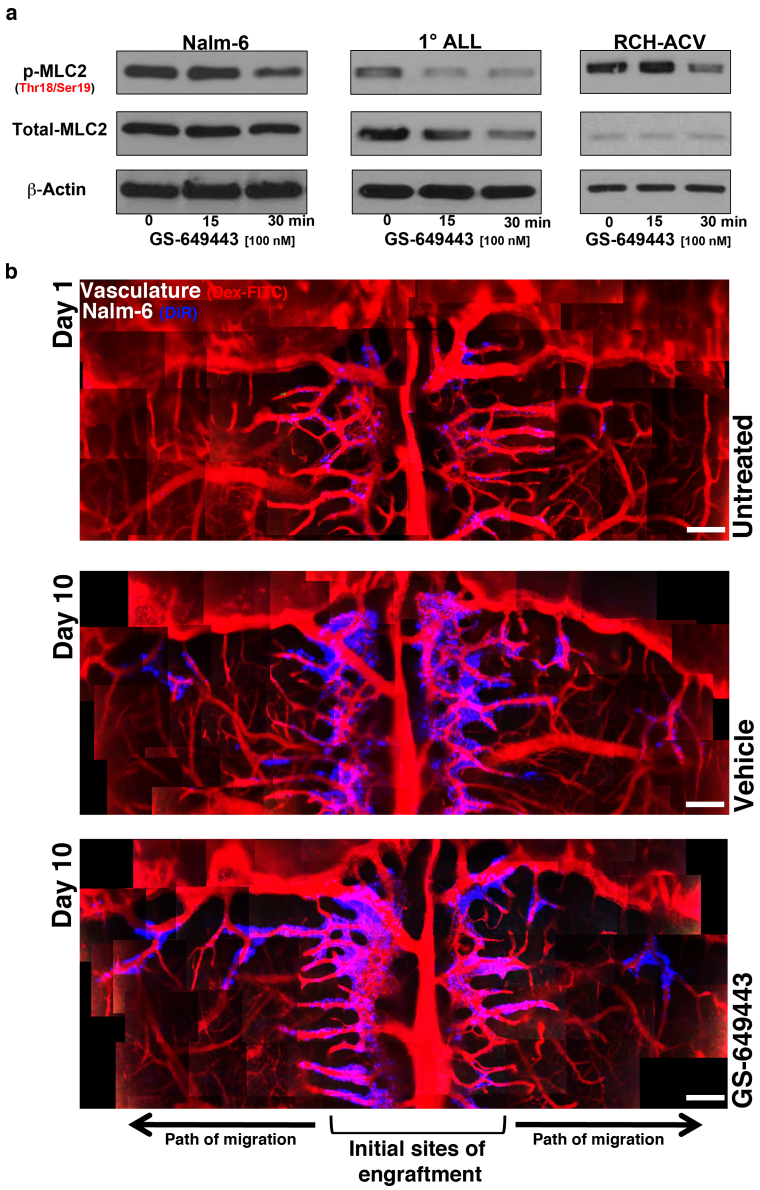
Extended Data Fig. 1 | PI3Kδ inhibition minimally decreases ALL peripheral disease burden. a, Representative haematoxylin and eosin (H&E)-stained histologic sections showing Nalm-6 disease burden in the femoral bone marrow and spleen of mice treated with vehicle or GS-649443, at matched time points. $n = 6$ mice per treatment group. **b**, Complete blood counts of vehicle- and GS-649443-treated mice, at matched time points. Paired two-tailed Student's *t*-test, $n = 5$ mice per treatment group, $P = 0.2158$; ns, not significant. **c**, Fold changes in CNS versus systemic disease burden in vehicle- and GS-649443-treated Nalm-6-engrafted mice. Data are mean ± s.e.m.; ANOVA with Tukey's multiple comparison test, $n = 5$ mice per treatment group, $P = 0.0314$. **d**, Representative histologic sections showing primary ALL disease burden in the femoral bone marrow and spleen of mice treated with vehicle or GS-649443 at matched time points. $n = 5$ mice per treatment group. **e**, Experimental plan for mice with primary ALL. **f**, Fold changes in CNS versus systemic disease burden in vehicle- and GS-649443-treated mice with primary ALL. Data are mean ± s.e.m.; ANOVA with Tukey, $n = 5$ mice per treatment group, $P = 0.00128$ (CNS versus bone marrow), $P = 0.0098$ (CNS versus spleen). **g**, Representative histologic sections showing RCH-ACV disease burden in the femoral bone marrow and spleen of mice treated with vehicle or GS-649443. $n = 4$ mice per treatment group. Arrowheads indicate RCH-ACV blasts. **h**, Kaplan–Meier survival curve for mice injected with RCH-ACV cells. Two-sided log rank Mantel-Cox, $n = 5$ mice per treatment group, $P = 0.0067$. **i**, Fold changes in CNS versus systemic disease burden in vehicle- and GS-649443-treated mice injected with RCH-ACV cells. Data are mean ± s.e.m.; ANOVA with Tukey, $n = 4$ mice per treatment group, $P = 0.0244$. Scale bars, 100 μm.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 2 | PI3Kδ inhibition at levels achievable in mouse CSF does not affect the apoptotic or cell cycle index of Nalm-6 cells; ROCK inhibition does not alter survival or disease burden in Nalm-6 leukaemic mice. a, Serum and brain tissue concentrations of GS-649443 in healthy SCID mice. Data are mean ± s.e.m.; $n = 5$ mice per group. **b**, Effect of PI3Kδ and Akt inhibition on transwell migration of primary ALL cells. Data are mean ± s.e.m.; ANOVA with Tukey, $n = 3$ technical replicates per group, $P = 0.5228$ (vehicle versus AZD5363), $P = 0.0549$ (vehicle versus idelalisib), $P = 0.0245$ (vehicle versus GS-649443). **c**, Percentage of annexin-V$^+$ Nalm-6 cells following in vitro treatment (72 h) with vehicle or GS-649443. Data are mean ± s.e.m.; unpaired two-sided Student's *t*-test, $n = 3$ biological replicates, $P = 0.2038$. **d**, Proportion of Nalm-6 cells in the G1, S or G2/M phase of the cell cycle following in vitro treatment (72 h) with vehicle or GS-649443. Data are mean ± s.e.m.; unpaired two-sided Student's *t*-test, $n = 3$ biological replicates, $P = 0.2216$ (G1), $P = 0.1405$ (S), $P = 0.0661$ (G2/M). **e**, Percentage of annexin-V$^+$ cells in the CD10$^+$ fraction of the CSF in vehicle- or GS-649443-treated mice. Paired two-sided Student's *t*-test, $n = 5$ mice per treatment group, $P = 0.6672$. **f**, Percentage of cells in G1 in the CD10$^+$ fraction of the CSF in vehicle- or GS-649443 treated mice. Paired two-sided Student's *t*-test, $n = 5$ mice per treatment group, $P = 0.1477$. **g**, Percentage cells in S or G2/M phase in the CD10$^+$ fraction of the CSF in vehicle- or GS-649443-treated mice. Paired two-sided Student's *t*-test, $n = 5$ mice per treatment group, $P = 0.5687$. **h**, Kaplan–Meier survival curves for Nalm-6-engrafted mice treated with or without the Rho-kinase (ROCK) inhibitor fasudil. Two-sided log rank Mantel–Cox, $P = 0.2843$; $n = 6$ per treatment group. **i**, Disease burden (CD10$^+$ cells) at end point in the bone marrow, spleen and CSF of vehicle- and fasudil-treated SCID mice engrafted with Nalm-6 cells. Data are mean ± s.e.m.; paired two-sided Student's *t*-test, $n = 4$ mice per treatment group, $P = 0.4534$ (bone marrow), $P = 0.3119$ (spleen), $P = 0.8026$ (CSF).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | PI3Kδ inhibition reduces myosin light chain activity but does not affect intra-bone marrow migration of Nalm-6 cells. a, Western blot of MLC2 levels in ALL cells following GS-649443 treatment. For gel source data, see Supplementary Fig. 1. Nalm-6 cells: $n = 3$ independent experiments; primary ALL cells: $n = 2$ independent experiments; RCH-ACV cells: $n = 3$ independent experiments. **b**, Representative intravital microscopy images of ALL-cell intra-bone marrow migration over time in vehicle- versus GS-649443-treated mice. $n = 3$ mice per treatment group. Scale bars, 250 μm.

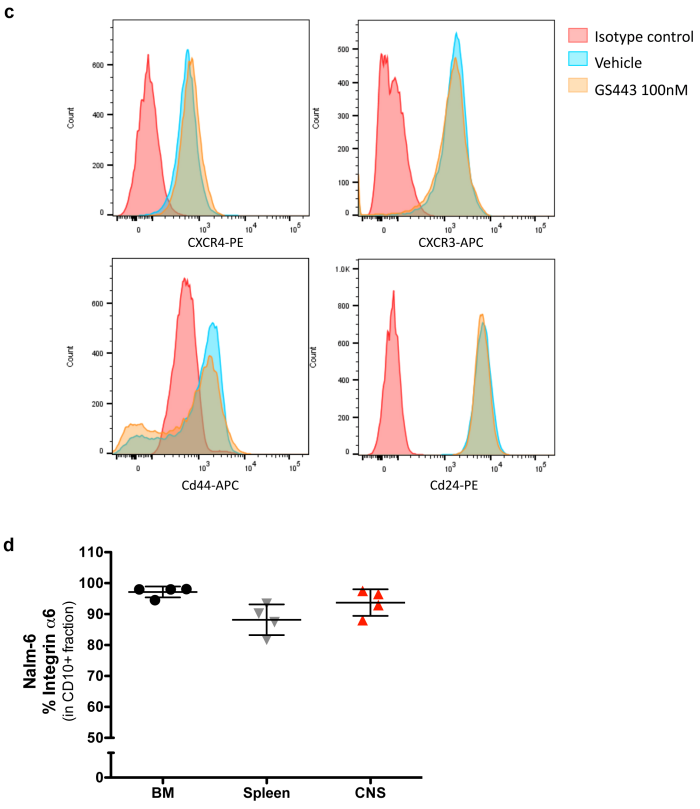
© 2018 Springer Nature Limited. All rights reserved.

a **BM array - vehicle vs. GS649443**

Gene Symbol	Fold Change	ANOVA p-value
SPP1	-3.30	0.018
ITGA6	**-1.91**	**0.013**
BCL2	-1.82	0.005
VAV3	-1.73	0.015
GNG7	-1.69	0.004
VASP	-1.69	0.022
SLC2A3	-1.57	0.037

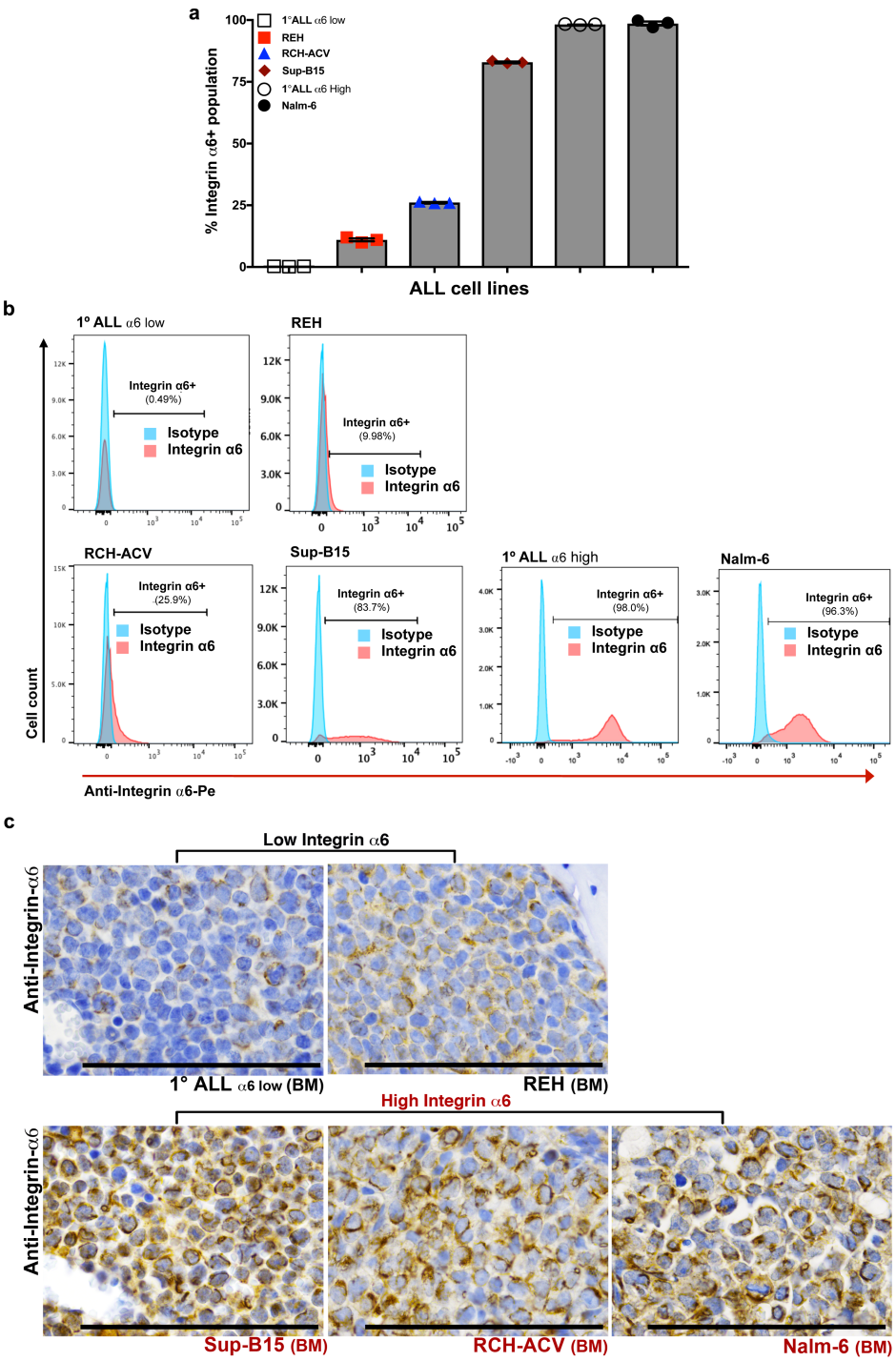
b **CSF array - vehicle vs. GS649443**

Gene Symbol	Fold Change	ANOVA p-value
IL1B	-3.35	0.019
SPP1	-3.33	0.001
SLC2A3	-2.98	0.008
NCF1	-2.85	0.012
ITGA6	**-2.51**	**0.015**
ADCY1	-2.18	0.013
VEGFA	-2.17	0.014
MAP2K1	-2.13	0.011
ATP2A3	-2.11	0.036
ATF5; MIR4751	-2.06	0.023
CALD1	-2.00	0.033
LPAR6	-1.98	0.007
ITPR3	-1.89	0.031
PIK3CG	-1.82	0.035
DDIT4	-1.80	0.020
RP11-380G5.2; PTEN	-1.73	0.042
LYN	-1.71	0.000
LAMA3	-1.70	0.002
MYLK2	**-1.64**	**0.004**
GNG7	-1.62	0.036
GRK6	-1.59	0.010
CALCA	-1.57	0.016
ITGAL	-1.54	0.007
VEGFB	-1.54	0.013
STAT1	-1.52	0.047



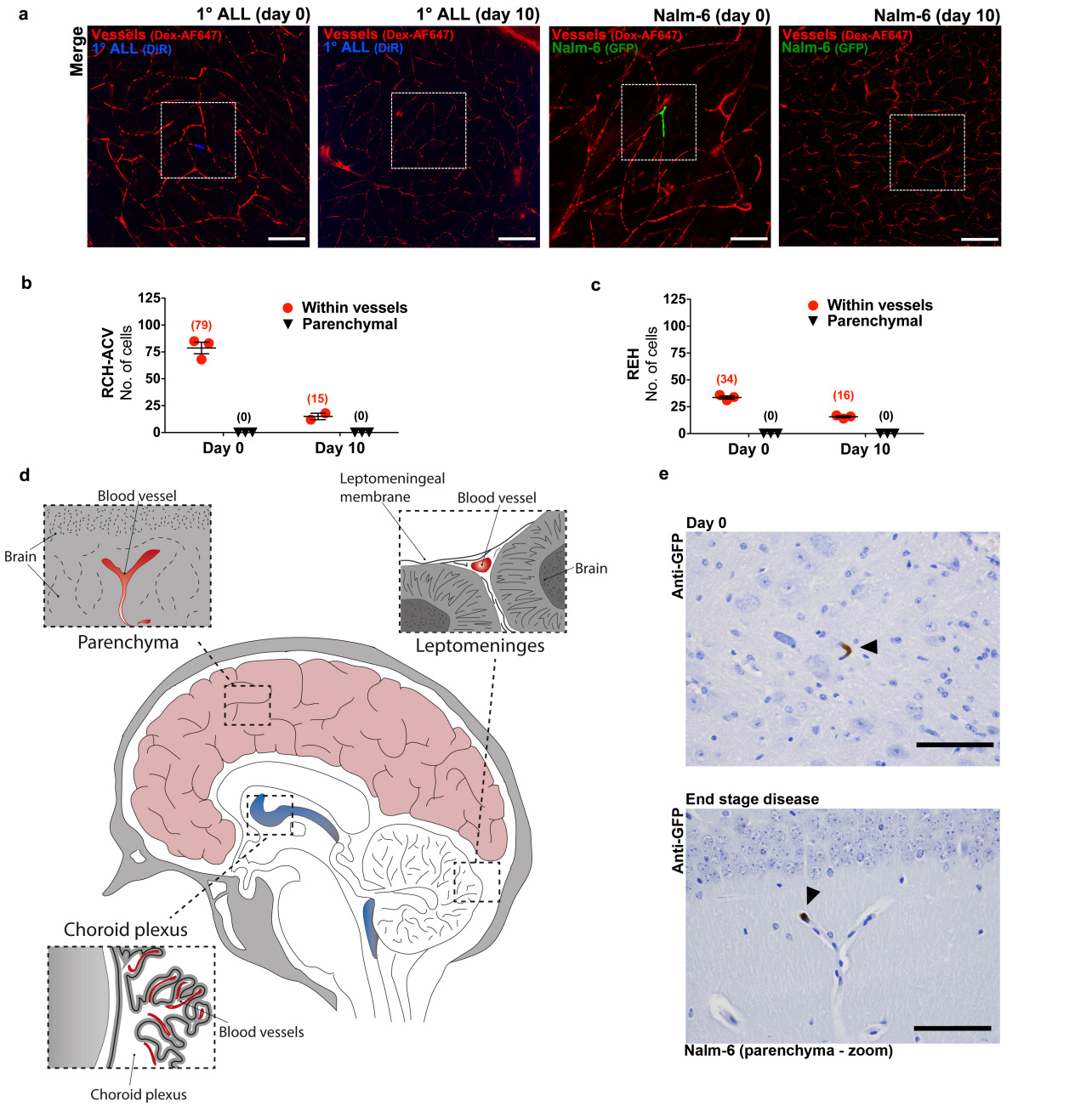
Extended Data Fig. 4 | Microarray analysis of focal adhesion and motility pathway genes in blasts from vehicle- versus GS-649443-treated mice and validation of key candidates. a, **b**, Differentially expressed focal adhesion and motility pathway genes identified by microarray analysis of blasts from vehicle- versus GS-649443-treated mice. $n = 6$ mice per treatment group. **c**, Microarray negative control validation. Representative flow cytometry analysis of Nalm-6 expression of CXCR4, CXCR3, CD44 and CD24 following treatment with vehicle or GS-649443. $n = 3$ independent experiments. **d**, ALL α6 integrin expression in vivo by flow cytometry. Data are mean ± s.e.m.; $n = 4$ mice.

© 2018 Springer Nature Limited. All rights reserved.



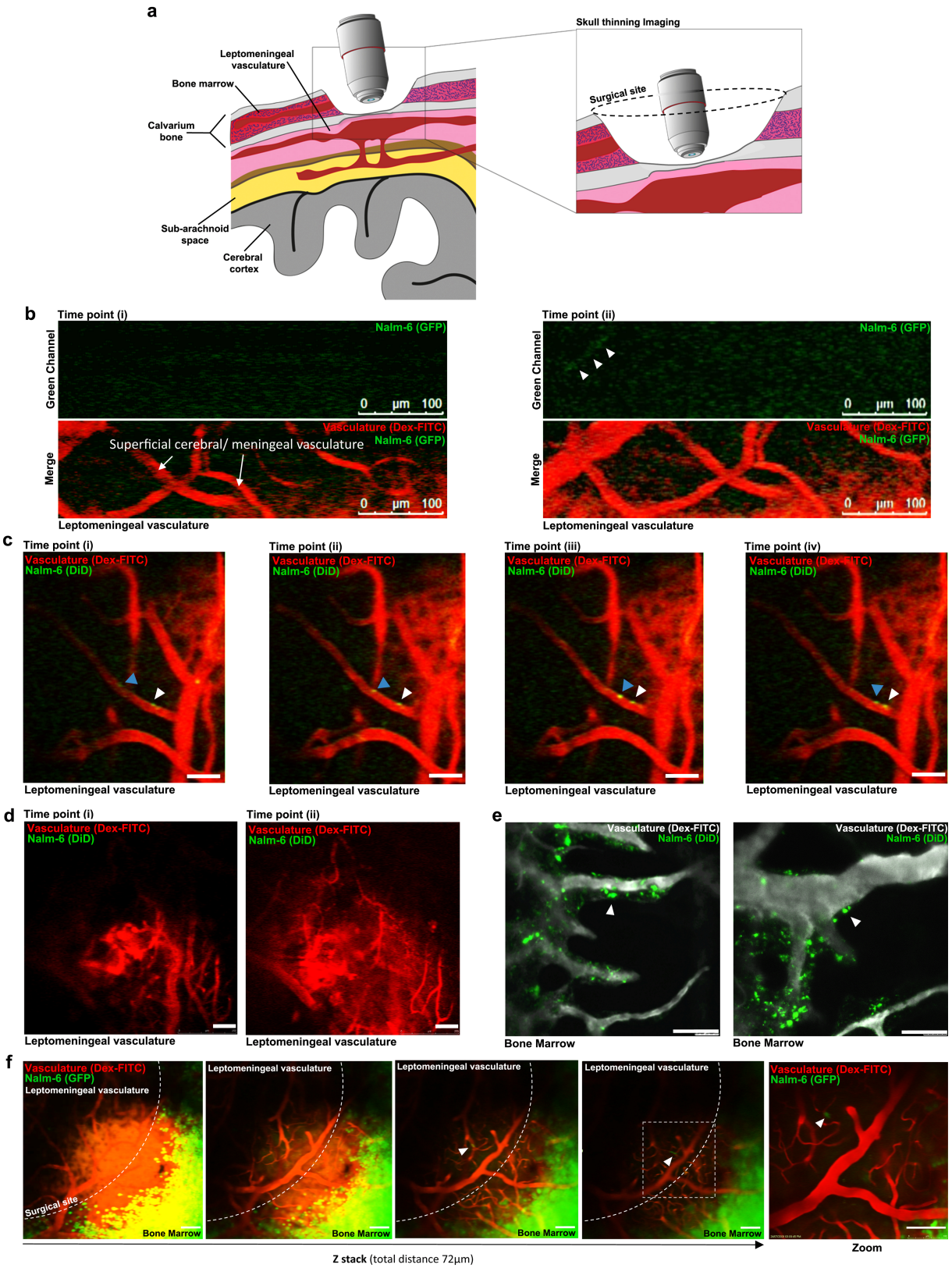
Extended Data Fig. 5 | Cell surface expression of α6 integrin is variable across a panel of ALL cell lines and primary human ALL cells in vitro and in vivo. a, Flow cytometry analysis of the percentage of α6 integrin$^+$ populations in ALL cell lines and primary ALL cells. Data are mean ± s.e.m.; $n = 3$ independent experiments per cell line. **b**, Representative flow cytometry histograms of α6 integrin expression in ALL cell lines and primary ALL cells. $n = 3$ independent experiments. **c**, Representative α6 integrin immunohistochemistry of femoral bone marrow from mice engrafted with low α6 integrin-expressing versus high α6 integrin-expressing ALL cells. α6 integrin low primary human ALL: $n = 6$ mice; REH: $n = 6$ mice; SUP-B15: $n = 3$ mice; RCH-ACV: $n = 6$ mice; Nalm-6: $n = 10$ mice.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | ALL cells do not invade through brain microvasculature. a, Representative confocal microscopy images of ALL cells located within brain microvessels at various time points after intravenous engraftment. $n = 3$ mice per cell line and time point, 12 micrographs per mouse. White boxes outline areas shown at higher magnification in Fig. 3a. **b**, **c**, Quantification of the number of ALL cells located in brain parenchymal tissue or within brain parenchymal microvessels at various time points after intravenous engraftment of RCH-ACV (**b**) and REH (**c**) ALL cells. Data are mean ± s.e.m.; $n = 3$ biological replicates per time point. **d**, Cartoon showing vasculature within the choroid plexus, leptomeninges and brain parenchyma in humans. **e**, Representative GFP immunohistochemistry staining of brain parenchyma of Nalm-6–GFP-engrafted mice on day 0 after intravenous engraftment ($n = 3$ mice) or at end stage disease ($n = 5$ mice). Images show close-ups of the ALL cells that are highlighted by arrowheads in Fig. 3f. Scale bars, 100 μm.

© 2018 Springer Nature Limited. All rights reserved.

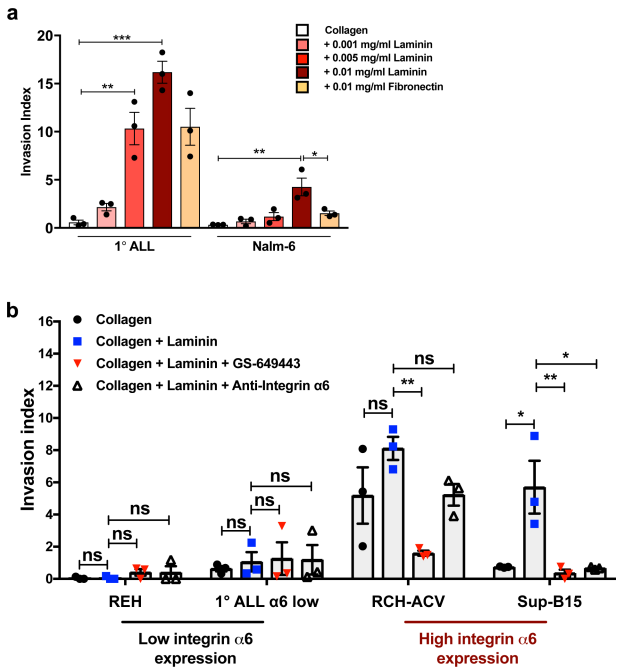


Extended Data Fig. 7 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

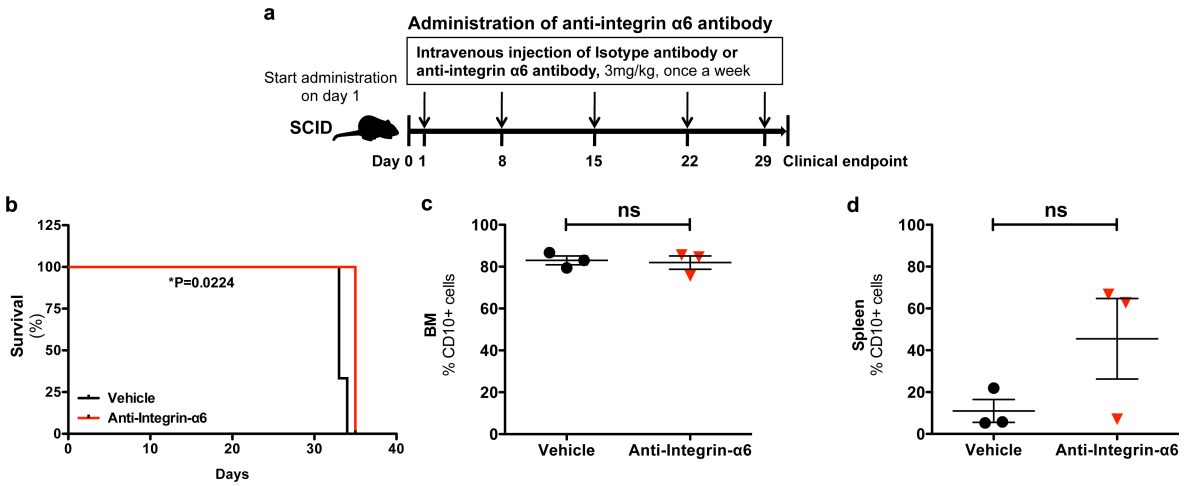
Extended Data Fig. 7 | Nalm-6 cells do not invade via diapedesis through the leptomeningeal blood–brain barrier. **a**, Graphic of thinned skull window and video-rate intravital confocal microscopy approach used to image the leptomeningeal vasculature at various time points after engraftment with ALL cells. **b**, Representative still images from video-rate intravital microscopy analysis of leptomeningeal and superficial cerebral vasculature (red) at 10 min after intravenous Nalm-6 engraftment (full video presented in Supplementary Videos 1, 2). Nalm-6 cells (green) are observed in circulation. $n = 12$ mice imaged on day 0 after engraftment. **c**, Series of still images of leptomeningeal and superficial cerebral vasculature at 45 min after intravenous Nalm-6 engraftment. A Nalm-6 cell is observed adherent to the luminal side of a leptomeningeal vessel (white arrowhead). A second Nalm-6 cell is observed rolling along the luminal wall of a leptomeningeal vessel (blue arrowhead) (see Supplementary Video 3). No invasion via diapedesis was observed during the entirety of each 2–4-h long imaging session on the day of engraftment. $n = 12$ mice imaged on day 0 after engraftment. **d**, Still images of the leptomeningeal and superficial cerebral vasculature 12 days after Nalm-6 engraftment. No Nalm-6 cells were observed in circulation or within the leptomeningeal tissue (see Supplementary Video 4; $n = 4$ mice, days 2, 4, 7 and 12 after engraftment). **e**, Representative images from intravital confocal microscopy of the calvarial bone marrow at 2 h after Nalm-6 engraftment. Numerous Nalm-6 cells (white arrowheads) are seen to have invaded through the bone marrow vasculature soon after intravenous engraftment. $n = 15$ mice. **f**, Series of still images of the z plane of the leptomeningeal and superficial cerebral vasculature of a leukaemic mouse engrafted with Nalm-6 cells at disease end point. $n = 7$ mice, days 37–39 after engraftment. Nalm-6 cells are observed in circulation (white arrowheads), but no cells are observed to invade (see Supplementary Video 5).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | ALL cell invasion along laminin matrices is regulated by PI3K signalling and α6 integrin. a, Primarly ALL and Nalm-6 in vitro invasion towards human CSF along collagen, collagen and laminin, or collagen and fibronectin matrices. Data are mean ± s.e.m.; ANOVA with Tukey, $n = 3$ biologically independent experiments; primary ALL: $P = 0.8937$ (collagen versus 0.001 mg ml^{-1} laminin), $P = 0.0020$ (collagen versus 0.005 mg ml^{-1} laminin), $P < 0.0001$ (collagen versus 0.01 mg ml^{-1} laminin), $P = 0.0604$ (0.01 mg ml^{-4} laminin versus fibronectin); Nalm-6: $P = 0.0012$ (collagen versus 0.01 mg ml^{-1} laminin), $P = 0.0150$ (0.01 mg ml^{-1} laminin versus fibronectin). **b**, Comparative in vitro invasion of low α6 integrin-expressing ALL cells versus high α6 integrin-expressing ALL cells towards human CSF along laminin matrices and effects of PI3Kδ inhibition or α6 integrin blockade. Data are mean ± s.e.m.; ANOVA with Tukey, $n = 3$ biologically independent experiments; REH: $P > 0.9999$ (collagen versus laminin), $P = 0.6967$ (collagen and laminin versus collagen and laminin and GS-649443), $P > 0.9999$ (collagen and laminin versus collagen and laminin and anti-integrin-α6); primary ALL: $P = 0.9758$ (collagen versus laminin), $P = 0.9974$ (collagen and laminin versus collagen and laminin and GS-649443), $P = 0.9993$ (collagen and laminin versus collagen and laminin and anti-integrin-α6); RCH: $P = 0.2446$ (collagen versus laminin), $P = 0.0079$ (collagen and laminin versus collagen and laminin and GS-649443), $P = 0.2549$ (collagen and laminin versus collagen and laminin and anti-integrin-α6); SUP-B15: $P = 0.0122$ (collagen versus laminin), $P = 0.0080$ (collagen and laminin versus collagen and laminin and GS-649443), $P = 0.0112$ (collagen and laminin versus collagen and laminin and anti-integrin-α6). *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Anti-α6 integrin-blocking antibody treatment prolongs survival, but does not alter disease burden in the bone marrow or spleen of Nalm-6-engrafted leukaemic mice. a, Schematic for the treatment of mice with the α6 integrin-neutralizing antibody. **b**, Kaplan–Meier survival curves for Nalm-6-engrafted mice treated with or without anti-integrin-α6 blocking antibodies. Two-sided log rank Mantel–Cox, $n = 3$ mice per treatment group, $P = 0.0224$. **c**, **d**, Disease burden at end point in the bone marrow and spleen of vehicle and anti-integrin-α6 antibody-treated Nalm-6-engrafted leukaemic mice. Data are mean ± s.e.m.; paired two-sided Student's *t*-test, $n = 3$ mice per treatment group, $P = 0.7874$ (bone marrow), $P = 0.1595$ (spleen).

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | α6 integrin expression is associated with CNS relapse independently of other known risk factors for CNS disease recurrence

Group	Patient #	Age at diagnosis	Gender	WBC (10^9/L) on presentation	Ph status	Other karyotypic abnormalities	Extra-medullary disease	CSF disease at diagnosis	CD56+ blasts	Relapse free survival (months)	Integrin α6 staining intensity by IHC
CNS relapse	1	2	male	44	neg	none	none	neg	pos	42.00	2
	2	19	male	40.2	pos	monosomy 7	none	neg	neg	15.00	2
	3	42	female	3.3	neg	unknown	none	neg	neg	14.00	1
	4	28	male	1.8	neg	none	none	neg	neg	20.00	3
	5	57	male	268	pos	none	none	neg	pos	7.00	0.5
	6	63	male	21.8	neg	Duplication 1q, gain of 8, deletion 9p and structural rearrangements of 2, 3, 6, 14 and 18	none	pos	neg	1.00	0
	7	74	male	13.3	neg	none	none	neg	neg	25.00	0
	8	45	female	5.1	pos	unknown	none	neg	neg	10.00	3
BM relapse	9	64	female	70.4	neg	none	none	neg	neg	28.00	0
	10	3	male	106.3	neg	none	testicular	neg	neg	54.00	0
	11	7	male	9.1	neg	Hyperdiploid clone with translocation 1;19, deletion 17p and multiple whole chromosome gains including 21, now with monosomy 7 in most cells	none	unknown	neg	9.00	0.5
	12	42	male	3.2	neg	none	none	neg	neg	24.00	0.5
	13	64	male	5.3	neg	3 copies of the BCR locus (22q11.2); 3 intact copies of the MLL locus (11q23)	none	neg	neg	12.00	2
	14	28	male	32.3	neg	Deletion of p16 (CDKN2A) at 9p21 and rearrangement of the IGH locus.	none	neg	neg	24.00	1
No relapse	15	31	male	4.7	pos	monosomy 7, additional material on the long (q) arm of chromosome 10 at 10q24 and the short (p) arm of chromosome 17 at 17p11.2	none	neg	neg	59.00	0.5
	16	6	male	3.4	neg	none	none	neg	neg	82.00	0
	17	29	male	18.4	neg	none	none	neg	pos	46.00	1
	18	2	female	11.2	neg	patient not found	none	neg	neg	89.00	0
	19	5	female	15.4	neg	Trisomy of chromosome 10, 17; Tetrasomy of chromosome 21	none	neg	pos	67.00	1
	20	21	male	1.2	neg	none	none	neg	pos	64.00	0.5
	21	3	male	18.6	neg	none	none	neg	pos	67.00	0
	22	31	male	4.2	neg	none	none	neg	neg	68.00	0
	23	9	male	12	neg	Hyperdiploidy 4, 10, 17, tetrasomy 21	none	neg	pos	72.00	0.5
	24	6	male	4.5	neg	Gain of chromosome 17 and 21 (gain of RUNX1)	none	neg	pos	72.00	0
	25	55	female	1.8	neg	none	none	neg	neg	64.00	0.5
	26	62	male	31.4	pos	BCR/ABL1 positive	none	neg	neg	103.00	0.5

Summary of clinical data for patients with ALL. Statistical testing was performed to determine potential associations between α6 integrin expression and known risk factors for CNS disease relapse. Fisher's exact test, $n=19$ patients without CNS relapse, $n=8$ patients with CNS relapse, α6 integrin immunohistochemistry score (0–0.5+, 1+ or 2+–3+) versus age at diagnosis (paediatric <15 years or adult): $P=0.8416$; immunohistochemistry score versus white blood cell count at diagnosis (low <30 (10^9 per l) or high >30 (10^9 per l), $P=0.8119$; immunohistochemistry score versus Philadelphia chromosome status: $P=0.4676$; α6 integrin immunohistochemistry score versus CD56 status: $P=0.2332$). BCR, breakpoint cluster region; BCR/ABL1, breakpoint cluster region/Abelson murine leukaemia viral oncogene homologue 1; BM, bone marrow; IgH, immunoglobulin heavy locus; IHC, immunohistochemistry; MLL, mixed-lineage leukaemia; Ph, Philadelphia chromosome; RUNX1, Runt-related transcription factor 1; WBC, white blood cells.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0367-9

Topological negative refraction of surface acoustic waves in a Weyl phononic crystal

Hailong He[1], Chunyin Qiu[1]*, Liping Ye[1], Xiangxi Cai[1], Xiying Fan[1], Manzhu Ke[1], Fan Zhang[2] & Zhengyou Liu[1,3]*

Reflection and refraction of waves occur at the interface between two different media. These two fundamental interfacial wave phenomena form the basis of fabricating various wave components, such as optical lenses. Classical refraction—now referred to as positive refraction—causes the transmitted wave to appear on the opposite side of the interface normal compared to the incident wave. By contrast, negative refraction results in the transmitted wave emerging on the same side of the interface normal. It has been observed in artificial materials[1–5], following its theoretical prediction[6], and has stimulated many applications including super-resolution imaging[7]. In general, reflection is inevitable during the refraction process, but this is often undesirable in designing wave functional devices. Here we report negative refraction of topological surface waves hosted by a Weyl phononic crystal—an acoustic analogue of the recently discovered Weyl semimetals[8–12]. The interfaces at which this topological negative refraction occurs are one-dimensional edges separating different facets of the crystal. By tailoring the surface terminations of the Weyl phononic crystal, constant-frequency contours of surface acoustic waves can be designed to produce negative refraction at certain interfaces, while positive refraction is realized at different interfaces within the same sample. In contrast to the more familiar behaviour of waves at interfaces, unwanted reflection can be prevented in our crystal, owing to the open nature of the constant-frequency contours, which is a hallmark of the topologically protected surface states in Weyl crystals[8–12].

Weyl semimetals[8–12], which feature doubly degenerate, linear band crossing points in three-dimensional momentum space, have become a research focus in the field of topological matter[13,14]. Weyl points are monopoles of Berry flux with topological charges characterized by quantized Chern numbers. In addition to the bulk properties that manifest the chiral anomaly of Weyl semimetals[9,15,16], their non-trivial band topology means that their boundaries host one-way gapless surface states[8,10,11]. In particular, the surface band dispersion at the Fermi energy can form open arcs that connect projected Weyl points of opposite topological charge. Weyl physics has recently been extended to artificial structures for classical waves, such as photonic and phononic crystals[17–25], in which materials with different optical or acoustic properties are arranged periodically. Soon after the experimental discovery of photonic Weyl points in a double-gyroid structure[18], the associated surface arc states were observed successfully[19–21,25]. In particular, more controllable structure design and less demanding signal detection have enabled macroscopic classical systems to be used[26–30] to explore the topological band physics originally proposed in electronic systems. Here, using a Weyl phononic crystal with precisely designed surface terminations, we present an experimental observation of topological negative refraction of the surface arc states. Our results advance the current knowledge of interfacial acoustics and enrich the Weyl physics in condensed-matter systems, providing a foundation for the exotic phenomena that arise from band topology.

In general, the response of an incident wave to a flat interface between two different sound media can be predicted using equifrequency contour (EFC) analysis, which is applicable to the interfacial systems that are formed both by naturally occurring homogenous media and by artificially constructed periodic structures[31]. As an extension of Snell's law for homogenous media, EFC analysis can be performed assuming three fundamental criteria: first, that the incoming and outgoing beams have equal frequency for the linear media considered here; second, that momentum is conserved parallel to the interface; and third, that the reflected and refracted beams must both leave the interface, yielding a constraint on their group velocities. In Fig. 1 we illustrate various sound responses at one- or two-dimensional interfaces that separate different sound media. In contrast to the normal, positive refraction (Fig. 1a) that is observed between two homogenous media (illustrated by EFCs expanding with frequency), anomalous negative refraction (Fig. 1b) emerges when the outgoing medium is replaced by a phononic crystal with a negative refraction index (illustrated by EFCs contracting with frequency). Note that all of these EFCs are closed orbits and inevitably produce unwanted interfacial reflection. Positive and negative refractions are both feasible for interfacial systems constructed using three-dimensional Weyl phononic crystals (Fig. 1c, d). By contrast, interfacial reflection can be made forbidden by taking advantage of the open EFCs of topological surface acoustic waves (SAWs), at which no reflected surface mode is allowed inherently. (The EFCs may also have a contribution from the bulk band projection of the Weyl phononic crystal, but for simplicity here we focus on the cases for which the surface arcs carry parallel momenta that are different from those of the bulk states; otherwise, possible energy leakage into the bulk must be taken into account.) As we show below, such interfacial phenomena can be realized in a single Weyl phononic crystal with precisely designed surface terminations, the interfaces of which are the one-dimensional edges shared by two adjacent facets.

We consider a 'woodpile' phononic crystal, which consists of stacked trilayer building blocks with broken inversion symmetry (Fig. 2a–c). Each trilayer unit consists of three identical square epoxy rods that are twisted anticlockwise by $2\pi/3$ along the z direction layer by layer, forming a triangular lattice in the x–y plane. The side length of each square rod is 9.8 mm and the in-plane and out-of-plane lattice constants are 29.4 mm. The experimental sample has a cuboid geometry with a size of 70.6 cm × 68.7 cm × 64.7 cm; that is, 22 trilayer units along the z direction and 24 × 27 structural units in the x–y plane. Termination geometries for the four side surfaces are prepared using a laser cutting technique. The side surfaces denoted YZ_1 and YZ_2 have the same appearance, whereas those denoted XZ_1 and XZ_2 are markedly different (Fig. 2b, d–f). Such a design gives the SAW EFCs the desired properties for realizing topological negative refraction (Fig. 1d) at one edge, together with positive refraction (Fig. 1c) at another edge of the same sample. As illustrated in Fig. 2d–f, we use three connected side surfaces (XZ_1, YZ_1 and XZ_2) to demonstrate the interfacial phenomena. To best characterize the one-way chiral SAWs and their interfacial responses, the three side surfaces are cast onto the same plane and a unified x axis pointing to the right is used for the surfaces XZ_1 and XZ_2. Acoustically

[1]Key Laboratory of Artificial Micro- and Nano-structures of Ministry of Education and School of Physics and Technology, Wuhan University, Wuhan, China. [2]Department of Physics, University of Texas at Dallas, Richardson, TX, USA. [3]Institute for Advanced Studies, Wuhan University, Wuhan, China. *e-mail: cyqiu@whu.edu.cn; zyliu@whu.edu.cn

© 2018 Springer Nature Limited. All rights reserved.

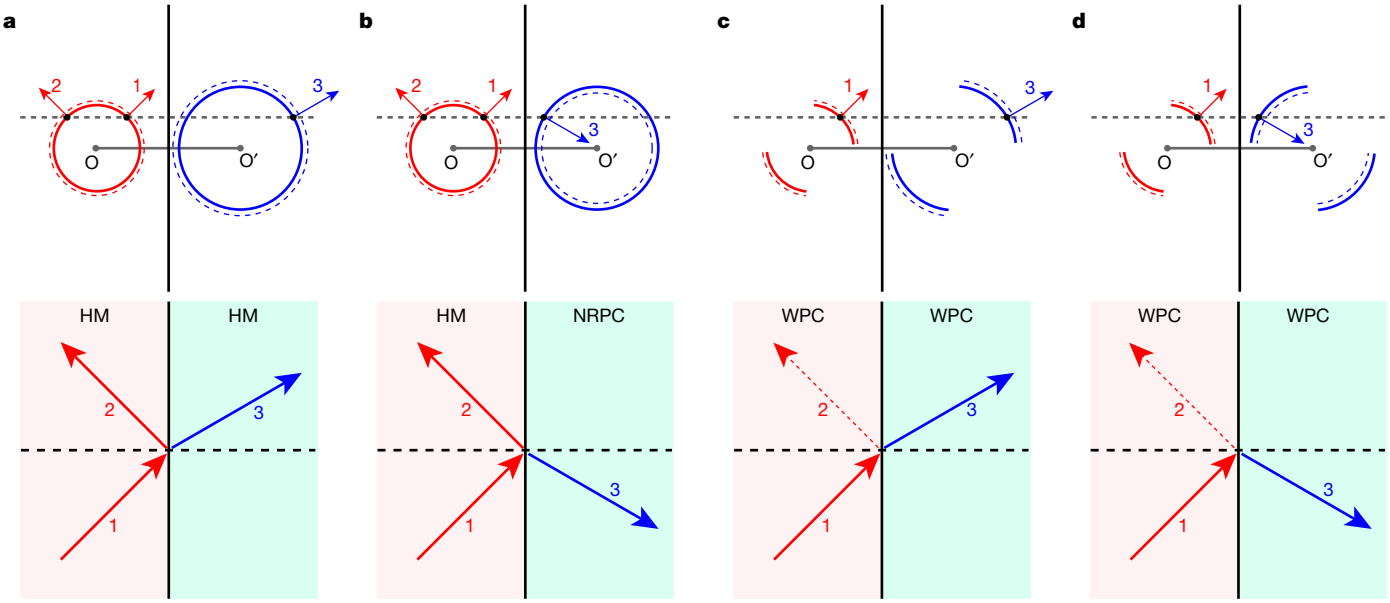


Fig. 1 | Schematics of different responses of sound waves to interfaces. The upper and lower panels are EFC analyses and associated beam propagations, respectively. In the upper panels, solid curves are EFCs at the operating frequency and dashed curves are EFCs at a slightly larger frequency. The arrows labelled 1, 2 and 3 point towards EFCs of increasing frequencies. For simplicity, circular EFCs (centred at O or O′), closed or open, are illustrated. In the lower panels, the arrows labelled 1, 2, and 3 indicate the incident, reflected and refracted beams, respectively, with group velocities defined by $\boldsymbol{v}_g = \nabla_{\boldsymbol{k}}\omega$, where $\boldsymbol{k}$ and ω are the wavevector and angular frequency, respectively. **a**, Conventional positive refraction at an interface between two different naturally occurring homogenous media (HM). **b**, Negative refraction when the outgoing medium in **a** is replaced by a phononic crystal designed with a negative refractive index (NRPC). **c**, Topological positive refraction of SAWs at an interface formed by two neighbouring facets of a Weyl phononic crystal (WPC). Two open arcs related by time-reversal symmetry are illustrated for each facet. **d**, The same as **c**, but for topological negative refraction. Interfacial reflection is forbidden in **c** and **d** (indicated by a dashed arrow) owing to the open EFCs, in contrast to the unavoidable reflection in **a** and **b** with closed EFCs.

hard epoxy plates, as trivial insulators for sound, are closely attached on the sample to host the topological surface arc states.

In Fig. 2g we show the bulk band dispersions for the Weyl phononic crystal along directions of high symmetry. This structurally simple phononic crystal has not only single Weyl points but also double Weyl points, which link together the lowest three bands (coloured lines). The single Weyl points, around which the dispersions are linear along all directions, emerge at K (or K′) and H (or H′) points, which are not invariant under time reversal. By contrast, the double Weyl points, around which the dispersions become quadratic in the k_x–k_y plane, are located at Γ and A points, as dictated by the time-reversal symmetry. The stability of these Weyl points at the high-symmetry momenta is a direct consequence of the three-fold screw symmetry of the Weyl phononic crystal. The topological charges of these Weyl points, as sources or sinks of Berry flux, can be calculated either by integrating Berry curvatures or by analysing the rotational eigenvalues at their high-symmetry momenta[22,32]. These topological charges and their distribution (Fig. 2g, h) lead to topologically non-trivial SAWs on the truncated surfaces parallel to the z direction (Extended Data Figs. 1, 2). This is demonstrated in Fig. 2i–k by the SAW dispersions (green lines) simulated at $k_z = 0.5\pi/h$ (where h is the lattice constant along the z direction) for the XZ_1, YZ_1 and XZ_2 surfaces, which traverse a wide gap between the lowest two bulk bands. Because all of these dispersions have positive slopes, the k_z-fixed SAWs propagate anticlockwise around the sample (viewed from the top). This one-way chiral behaviour of the SAWs can be understood by the fact that the spiral interlayer coupling in the phononic crystal introduces a synthetic gauge flux that threads through the sample along the z direction. The surface conditions greatly influence the properties of the SAW dispersions that are hosted by the XZ_1 and XZ_2 surfaces (Fig. 2i, k). This is further demonstrated in Fig. 2l, n by the open EFCs (green lines) simulated at 5.75 kHz, the frequency of the Weyl points with topological charge +1. Specifically, for $k_z > 0$ the SAWs at the XZ_2 surface always carry a downward component of the group velocity, in contrast to the upward component at the XZ_1 surface. (For the SAWs with $k_z < 0$, the group velocities reverse automatically owing to time-reversal symmetry.) Such contrasting features contribute substantially to the distinct SAW responses to the two one-dimensional edges of the YZ_1 surface (shared with the XZ_1 or XZ_2 surfaces). In addition, these surface EFCs exhibit gentle curvature for a wide range of momenta and the majority of SAWs propagate towards nearly the same direction. This facilitates the experimental observation of positive and negative refraction of topologically protected SAWs.

The presence of the topologically non-trivial SAWs is validated by our experiments with airborne sound. A broadband point-like sound generator is positioned under the cover plate to excite the SAWs and a subwavelength-sized sound probe is inserted inside the sample to scan the surface-pressure signal point by point. In particular, in each sample surface, the sound source is placed on a specific surface corner (Fig. 2d–f) to stimulate the one-way chiral SAWs. In this way, all SAWs that propagate rightwards with $k_z > 0$ are selectively excited. This not only reduces the finite-size effect but also silences the information from the time-reversal counterparts (that is, the SAWs with $k_z < 0$). Using a Fourier transformation of the near-field pressure distributions, we map out the SAW dispersions in the surface Brillouin zones. In Fig. 2i–k we present the experimental SAW dispersions (colour scale) in the k_x–k_y plane for $k_z = 0.5\pi/h$; Fig. 2l–n displays the measured surface EFCs at the Weyl frequency of 5.75 kHz. Excellent agreement is found between the experiments (colour scale) and simulations (green lines) for all three surfaces, despite band broadening due to the finite-size effect. The invisible excitation of the surface arc state with $k_z < 0$ also reflects negligible scattering from the well-excited surface arc state with $k_z > 0$.

We now turn to the interfacial responses of the topologically protected SAWs. We carried out two comparative experiments, one for the interface between the XZ_1 and YZ_1 surfaces, and the other for the interface between the YZ_1 and XZ_2 surfaces. In Fig. 3 we show the measured near-field pressure distributions at 5.75 kHz. In each case, the point-like sound source is positioned at the bottom of the initial surface (XZ_1 or YZ_1) to excite the SAWs with $k_z > 0$, which propagate towards the upper right of the surface. As anticipated from the EFC analysis (inset in Fig. 3a), Fig. 3a displays a typical positive refraction

© 2018 Springer Nature Limited. All rights reserved.

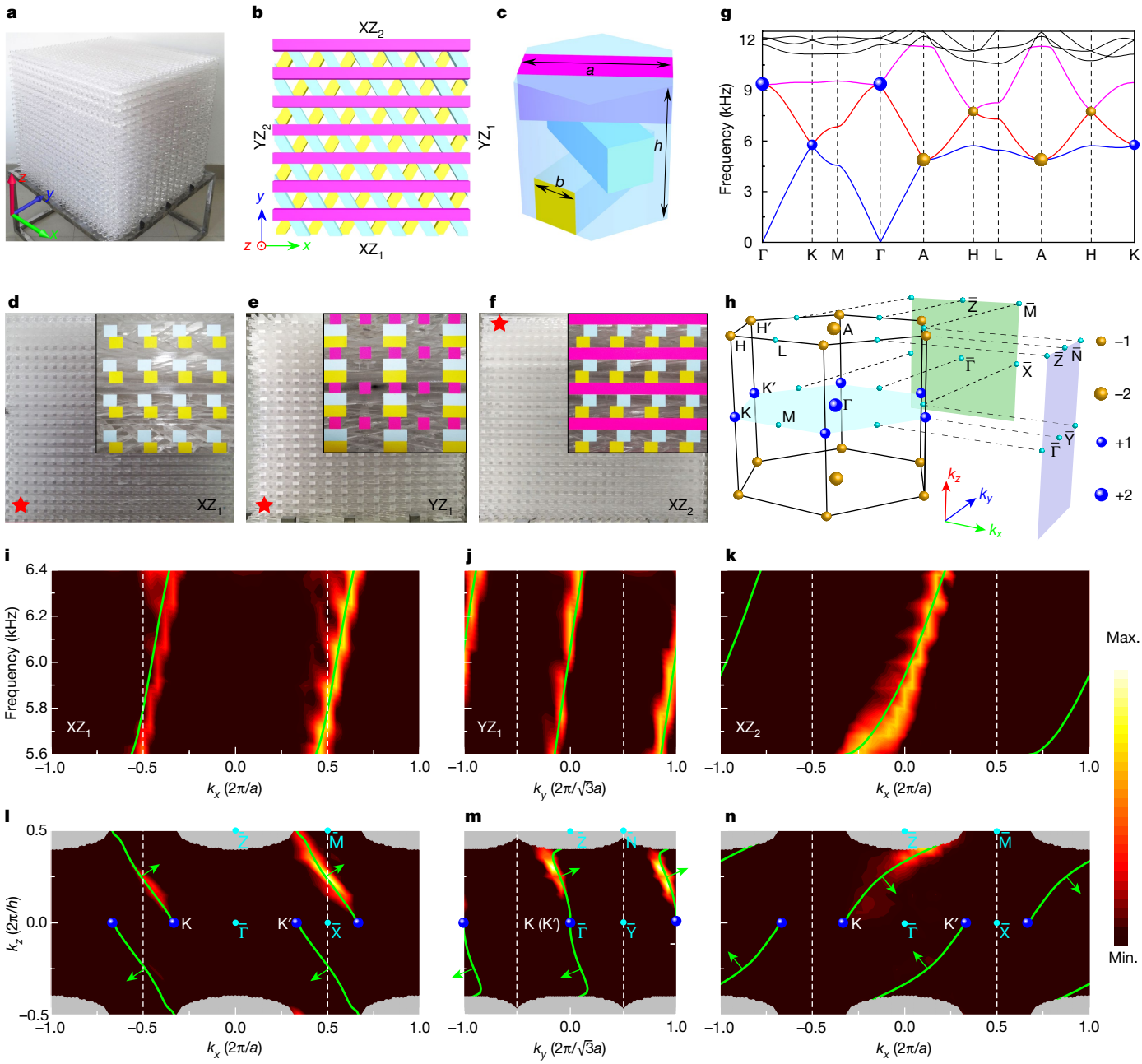


Fig. 2 | The Weyl phononic crystal and topologically protected SAWs. **a**, An image of the experimental sample. **b**, Schematic top view of the trilayer-based sample. XZ_1, YZ_1, XZ_2 and YZ_2 label the four side surfaces. **c**, Geometry of the unit cell, with $a = h = 3b = 29.4$ mm. **d**–**f**, Front views of the three surfaces XZ_1, YZ_1, and XZ_2, respectively. At each surface, the red star denotes the position of a point-like sound source for experimentally generating one-way chiral SAWs and the coloured segments in the insets indicate the fine structures of the surface termination. **g**, Bulk band dispersions simulated along high-symmetry directions. The coloured lines represent the lowest three bands. **h**, The first bulk Brillouin zone of the Weyl phononic crystal and associated projected surface Brillouin zones. The coloured spheres in **g** and **h** label Weyl points with different topological charges. **i**–**k**, Simulated SAW dispersions (green lines) at $k_z = 0.5\pi/h$ for the three side surfaces XZ_1, YZ_1 and XZ_2, respectively, agree very well with our measurements (bright colours in the colour scale, which represents the Fourier transformation of the measured pressure field). **l**–**n**, The corresponding EFCs in the extended surface Brillouin zones, simulated and measured at the Weyl frequency of 5.75 kHz. The grey regions display the projected bulk bands, the blue spheres label the projected Weyl points K and K′, and the green arrows indicate the directions of the SAW group velocities.

effect when the sound signal traverses across the interface to reach the second surface YZ_1, where the refracted beam travels upwards and deflects on the opposite side of the interface normal (white dashed line) with respect to the incoming beam. The finite widths of the incoming and outgoing beams in real space stem mostly from the finite lengths of the open EFCs with $k_z > 0$. By contrast, Fig. 3b shows negative refraction, with the outgoing beam propagating downwards and deflecting into the same side of the interface normal. This is consistent with the fact that the z component of the group velocity is reversed for the SAWs travelling from the YZ_1 to the XZ_2 surface (inset in Fig. 3b). For both cases, the measured incident and refracted angles agree well with predictions from the simulated and measured EFCs (Fig. 2l–n); these angles give estimated refraction indices of 0.6 and −2.2 for the cases of positive and negative refraction, respectively. In particular, for each interfacial system, no clear reflection signal is observed near the interface, unlike the conventional wave response to an interface. The reflection immunity to the time-reversal-related surface arc states is further confirmed by the Fourier transforms of the pressure fields on

© 2018 Springer Nature Limited. All rights reserved.

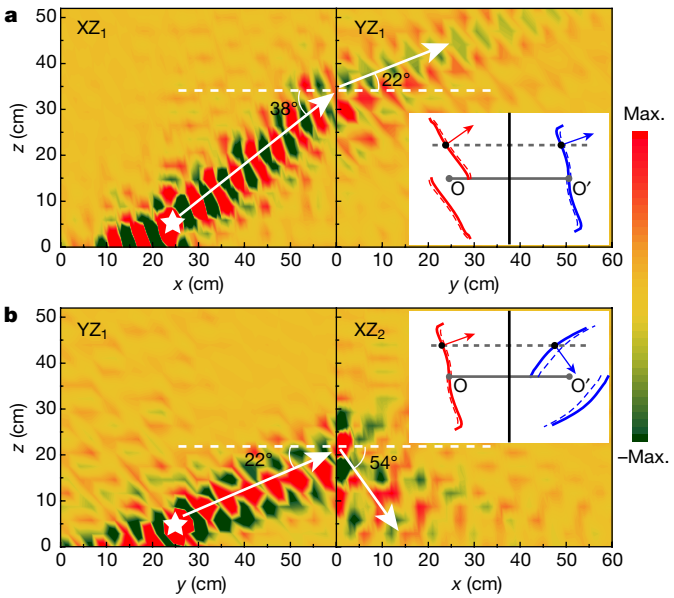


Fig. 3 | Experimental observation of topological negative refraction. **a**, **b**, Positive (**a**) and negative (**b**) responses of SAWs to the two one-dimensional edges of the YZ_1 surface (shared with the XZ_1 and XZ_2 surfaces, respectively). The near-field pressure distributions (colour scale) are measured at 5.75 kHz. The white stars label the positions of the point-like sound sources and the white arrows indicate the directions of the propagating beams. Insets, schematics of the echo-free interfacial phenomena, illustrated according to the EFC analysis in Fig. 1. The solid and dashed curves depict the simulated EFCs at 5.75 kHz and 5.85 kHz, respectively. In both cases, the experimentally measured beam propagations agree well with the EFC analysis and exhibit negligible interfacial reflections.

the incident surfaces. Physically, such interfacial phenomena benefit from the open nature of the surface arcs of a Weyl crystal. The drop in pressure amplitude at the interface is probably caused by the wave scattering into the bulk states and the mismatch of SAW properties between different surfaces. The signal reduction as the beam propagates stems mostly from unavoidable material absorption, broadening of the finite-width beam and scattering into bulk states. In addition to fabricating a cleaner sample, improvements can be achieved by injecting a sound source with a narrower range of k_z that is not carried by any projected bulk state.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0367-9

Received: 14 November 2017; Accepted: 5 June 2018;
Published online 1 August 2018.

1. Shelby, R. A., Smith, D. R. & Schultz, S. Experimental verification of a negative index of refraction. *Science* **292**, 77–79 (2001).
2. Cubukcu, E., Aydin, K., Ozbay, E., Foteinopoulou, S. & Soukoulis, C. M. Negative refraction by photonic crystals. *Nature* **423**, 604–605 (2003).
3. Yao, J. et al. Optical negative refraction in bulk metamaterials of nanowires. *Science* **321**, 930 (2008).
4. Yang, S. et al. Focusing of sound in a 3D phononic crystal. *Phys. Rev. Lett.* **93**, 024301 (2004).
5. Zhang, S., Yin, L. & Fang, N. Focusing ultrasound with an acoustic metamaterial network. *Phys. Rev. Lett.* **102**, 194301 (2009).
6. Veselago, V. G. The electrodynamics of substances with simultaneously negative values of ε and μ. *Sov. Phys. Usp.* **10**, 509 (1968).
7. Pendry, J. B. Negative refraction makes a perfect lens. *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
8. Wan, X., Turner, A. M., Vishwanath, A. & Savrasov, S. Y. Topological semimetal and Fermi-arc surface states in the electronic structure of pyrochlore iridates. *Phys. Rev. B* **83**, 205101 (2011).
9. Burkov, A. A. & Balents, L. Weyl semimetal in a topological insulator multilayer. *Phys. Rev. Lett.* **107**, 127205 (2011).
10. Xu, S.-Y. et al. Discovery of a Weyl fermion semimetal and topological Fermi arcs. *Science* **349**, 613–617 (2015).
11. Lv, B. Q. et al. Experimental discovery of Weyl semimetal TaAs. *Phys. Rev. X* **5**, 031013 (2015).
12. Soluyanov, A. A. et al. Type-II Weyl semimetals. *Nature* **527**, 495–498 (2015).
13. Hasan, M. Z. & Kane, C. L. Topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
14. Qi, X.-L. & Zhang, S.-C. Topological insulators and superconductors. *Rev. Mod. Phys.* **83**, 1057–1110 (2011).
15. Xu, G., Weng, H., Wang, Z., Dai, X. & Fang, Z. Chern semimetal and the quantized anomalous hall effect in $HgCr_2Se_4$. *Phys. Rev. Lett.* **107**, 186806 (2011).
16. Huang, X. et al. Observation of the chiral-anomaly-induced negative magnetoresistance in 3D Weyl semimetal TaAs. *Phys. Rev. X* **5**, 031023 (2015).
17. Lu, L., Fu, L., Joannopoulos, J. D. & Soljacic, M. Weyl points and line nodes in gyroid photonic crystals. *Nat. Photon.* **7**, 294–299 (2013).
18. Lu, L. et al. Experimental observation of Weyl points. *Science* **349**, 622–624 (2015).
19. Chen, W. J., Xiao, M. & Chan, C. T. Photonic crystals possessing multiple Weyl points and the experimental observation of robust surface states. *Nat. Commun.* **7**, 13038 (2016).
20. Noh, J. et al. Experimental observation of optical Weyl points and Fermi arc-like surface states. *Nat. Phys.* **13**, 611–617 (2017).
21. Yang, B. et al. Direct observation of topological surface-state arcs in photonic metamaterials. *Nat. Commun.* **8**, 97 (2017).
22. Chang, M. L., Xiao, M., Chen, W. J. & Chan, C. T. Multiple Weyl points and the sign change of their topological charges in woodpile photonic crystals. *Phys. Rev. B* **95**, 125136 (2017).
23. Xiao, M., Chen, W. J., He, W. Y. & Chan, C. T. Synthetic gauge flux and Weyl points in acoustic systems. *Nat. Phys.* **11**, 920–924 (2015).
24. Yang, Z. & Zhang, B. Acoustic type-II Weyl nodes from stacking dimerized chains. *Phys. Rev. Lett.* **117**, 224301 (2016).
25. Li, F., Huang, X., Lu, J., Ma, J. & Liu, Z. Weyl points and Fermi arcs in a chiral phononic crystal. *Nat. Phys.* **14**, 30–34 (2018).
26. Wang, Z., Chong, Y., Joannopoulos, J. D. & Soljacic, M. Observation of unidirectional backscattering-immune topological electromagnetic states. *Nature* **461**, 772–775 (2009).
27. Rechtsman, M. C. et al. Photonic floquet topological insulators. *Nature* **496**, 196–200 (2013).
28. Susstrunk, R. & Huber, S. D. Observation of phononic helical edge states in a mechanical topological insulator. *Science* **349**, 47–50 (2015).
29. Lu, J. et al. Observation of topological valley transport of sound in sonic crystals. *Nat. Phys.* **13**, 369–374 (2017).
30. Gao, F. et al. Topologically protected refraction of robust kink states in valley photonic crystals. *Nat. Phys.* **14**, 140–144 (2018).
31. Notomi, M. Theory of light propagation in strongly modulated photonic crystals: refractionlike behavior in the vicinity of the photonic band gap. *Phys. Rev. B* **62**, 10696–10705 (2000).
32. Fang, C., Gilbert, M. J., Dai, X. & Bernevig, B. A. Multi-Weyl topological semimetals stabilized by point group symmetry. *Phys. Rev. Lett.* **108**, 266802 (2012).

Acknowledgements We thank M. Xiao for discussions. This work is supported by the National Basic Research Program of China (grant number 2015CB755500), the National Natural Science Foundation of China (grant numbers 11774275, 11674250, 11534013 and 11747310) and the Natural Science Foundation of Hubei Province (grant number 2017CFA042). F.Z. was supported by UT Dallas research enhancement funds.

Reviewer information *Nature* thanks A. Grushin, B. Zhang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.Q. and Z.L. conceived the idea and supervised the project. H.H. performed the simulations. H.H., L.Y., X.F. and M.K. carried out the experiments. C.Q., H.H., X.C., F.Z. and Z.L. analysed the data and wrote the manuscript. All authors contributed to scientific discussions of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0367-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.Q. and Z.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Numerical simulations. All full-wave simulations are performed using a commercial solver package (COMSOL Multiphysics). The epoxy material used in our experiments is safely modelled as an acoustically rigid material, given the great impedance mismatch with respect to the air background. The speed of sound in air of 346 m s^{-1} is used for our experiment at room temperature (23 °C). The bulk band dispersions (Fig. 2g) are calculated for a single unit cell (Fig. 2c) with Bloch boundary conditions in all three-dimensional directions. To calculate the k_z-fixed surface band dispersions (Fig. 2i–k), we consider infinitely large slab structures with terminations and thicknesses specified in the text. In each case, the slab is thick enough to avoid the coupling between the SAWs hosted by the two different surfaces. Rigid boundary conditions are applied to both slab surfaces, whereas Bloch boundary conditions are used for the remaining two directions, with fixed k_z and varying k_x or k_y. In addition to the projected bulk bands, each numerical configuration gives the SAWs for both surfaces simultaneously, which can be further distinguished by inspecting the surface localization of the field distributions. Similarly, the EFCs (Fig. 2l–n) are extracted by scanning the surface Brillouin zone at 5.75 kHz. In addition to the numerical data provided in Fig. 2, simulations of the bulk and surface dispersions are provided in Extended Data Figs. 1 and 3.

Experimental measurements. To excite the SAWs experimentally, a broadband sound signal is launched from a narrow tube (radius of about 4.0 mm) that penetrates the epoxy plate covering the Weyl phononic crystal. The sound source behaves like a point source for the wavelength of about 60 mm used here. The surface field is scanned point by point through a microphone (radius of about 3.5 mm, B&K Type 4187) that is inserted inside the sample. The field confinement of the topologically protected SAWs is checked by detecting the pressure distributions away from the surface (Extended Data Fig. 4). The scanning steps are 5.0 mm, 25.4 mm and 29.4 mm along the x, y and z directions, respectively. Both the amplitude and phase of the local pressure field can be recorded and frequency-resolved by using a multi-analyser system (B&K Type 3560B). To map out the EFCs (Fig. 2l–n) in the surface Brillouin zone, a two-dimensional Fourier transformation is performed for the measured spatial pressure distributions at a given frequency. This also gives the frequency-dependent dispersion curves for fixed k_z (Fig. 2i–k).

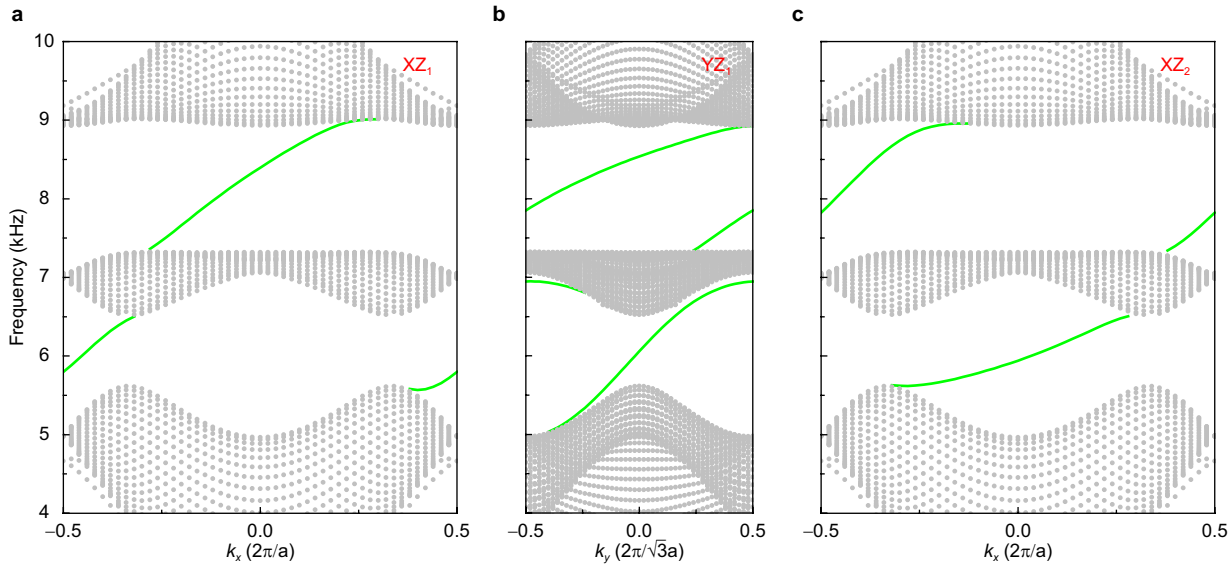
Discussions on our sample design. Here we provide a brief discussion on our woodpile Weyl phononic crystal, which is structurally simple, easily fabricated and convenient for field scanning in our sound experiments in air. This Weyl phononic crystal system, which relies on the three-fold screw symmetry and is beyond the description of tight-binding model[22], exhibits a clean bulk band structure and a broad topologically non-trivial frequency window (Extended Data Fig. 1). As a consequence of the crystalline symmetry, the single and double Weyl points are both located at high-symmetry points that are well separated in momentum space. The three-fold screw symmetry dictates that the single Weyl points stay at the non-time-reversal-invariant high-symmetry points, such as K (or K′) and H (or H′), and that the double Weyl points stay at the time-reversal-invariant high-symmetry points, such as Γ and A. Such symmetry requirements also provide extra protection of the Weyl points against pairwise annihilation. These merits (such as a wide energy window, fixed momenta of Weyl points and long distances between them) greatly facilitate numerical and experimental searches for Weyl points, and the confirmation of surface arc states and associated interfacial phenomena.

In addition, the single or double Weyl points of opposite charge sit at different frequencies (see Fig. 2g), owing to the chiral nature of the phononic crystal structure. This is very different from the Weyl semimetal TaAs[10,11], in which the Weyl points of opposite sign emerge at the same energy because of the mirror symmetry. In solid systems, chiral Weyl semimetals that lack mirror symmetries are very attractive because they are predicted to host a quantized circular photogalvanic effect[33]. Efforts have been devoted to exploring Weyl semimetals with chiral structures[33–37]. These include the so-called Kramers Weyl semimetals[34,35], in which the Weyl points emerge stably at the time-reversal-invariant points. (This mechanism requires strong spin–orbit coupling and therefore differs from our acoustic system, which is essentially spinless.) In trigonal tellurium and selenium[34], the Kramers-theorem-enforced Weyl semimetal phase takes place only under additional pressure. Recently, more material candidates for Kramers Weyl semimetals have been revealed in a theoretical study[35]. Other structurally chiral solid systems have also been proposed, such as CoSi[36] and $SrSi_2$[37], but have complex band structures[36] or are difficult to grow as single crystals[37]. The design route of our Weyl phononic crystal could potentially serve as a reference for further searches for chiral Weyl crystals in solid-state systems.

Data availability. The data that support the findings of this study are available from the corresponding authors on reasonable request.

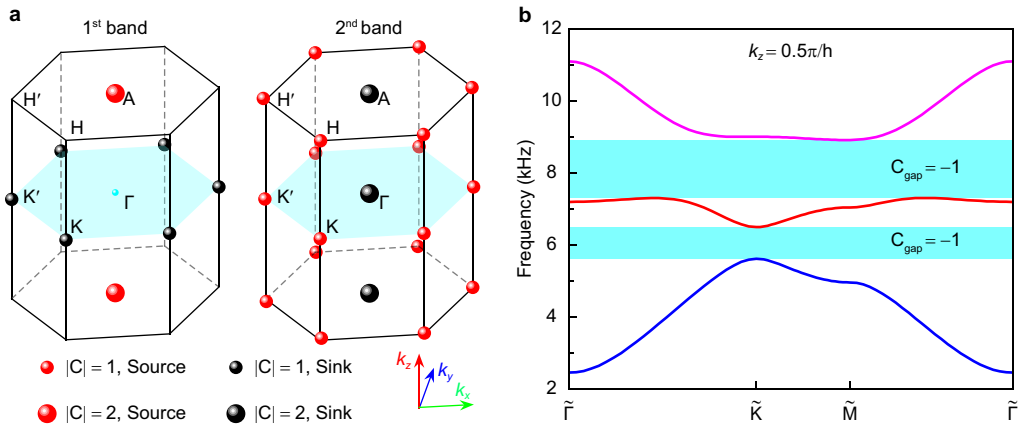
33. de Juan, F., Grushin, A. G., Morimoto, T. & Moore, J. E. Quantized circular photogalvanic effect in Weyl semimetals. *Nat. Commun.* **8**, 15995 (2017).
34. Hirayama, M., Okugawa, R., Ishibashi, S., Murakami, S. & Miyake, T. Weyl node and spin texture in trigonal tellurium and selenium. *Phys. Rev. Lett.* **114**, 206401 (2015).
35. Chang, G. et al. Universal topological electronic properties of nonmagnetic chiral crystals. *Preprint at* https://arXiv.org/abs/1611.07925 (2018).
36. Chang, G. et al. Unconventional chiral fermions and large topological Fermi arcs in RhSi. *Phys. Rev. Lett.* **119**, 206401 (2017).
37. Huang, S. et al. New type of Weyl semimetal with quadratic double Weyl fermions. *Proc. Natl Acad. Sci. USA* **113**, 1180–1185 (2016).

© 2018 Springer Nature Limited. All rights reserved.



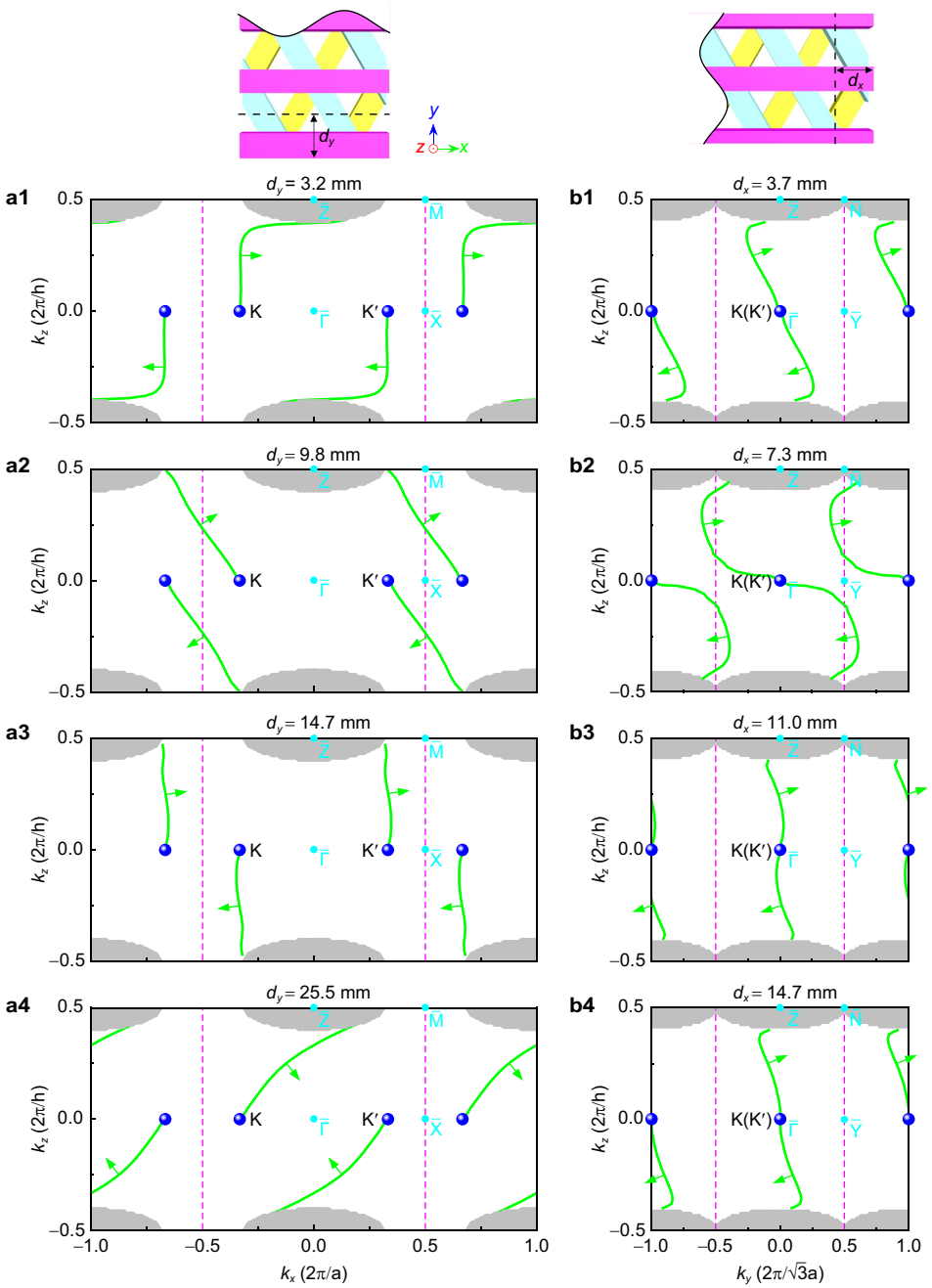
Extended Data Fig. 1 | Simulated surface band dispersions in a wide frequency range. a–c, The data are evaluated at $k_z = 0.5\pi/h$ for the three side surfaces XZ_1, YZ_1 and XZ_2, respectively, the terminations of which are specified in the main text. Two gapless surface bands (green lines) traverse the two gaps between the lowest three projected bulk bands (grey regions). We focus on the lowest surface band in the main text.

© 2018 Springer Nature Limited. All rights reserved.



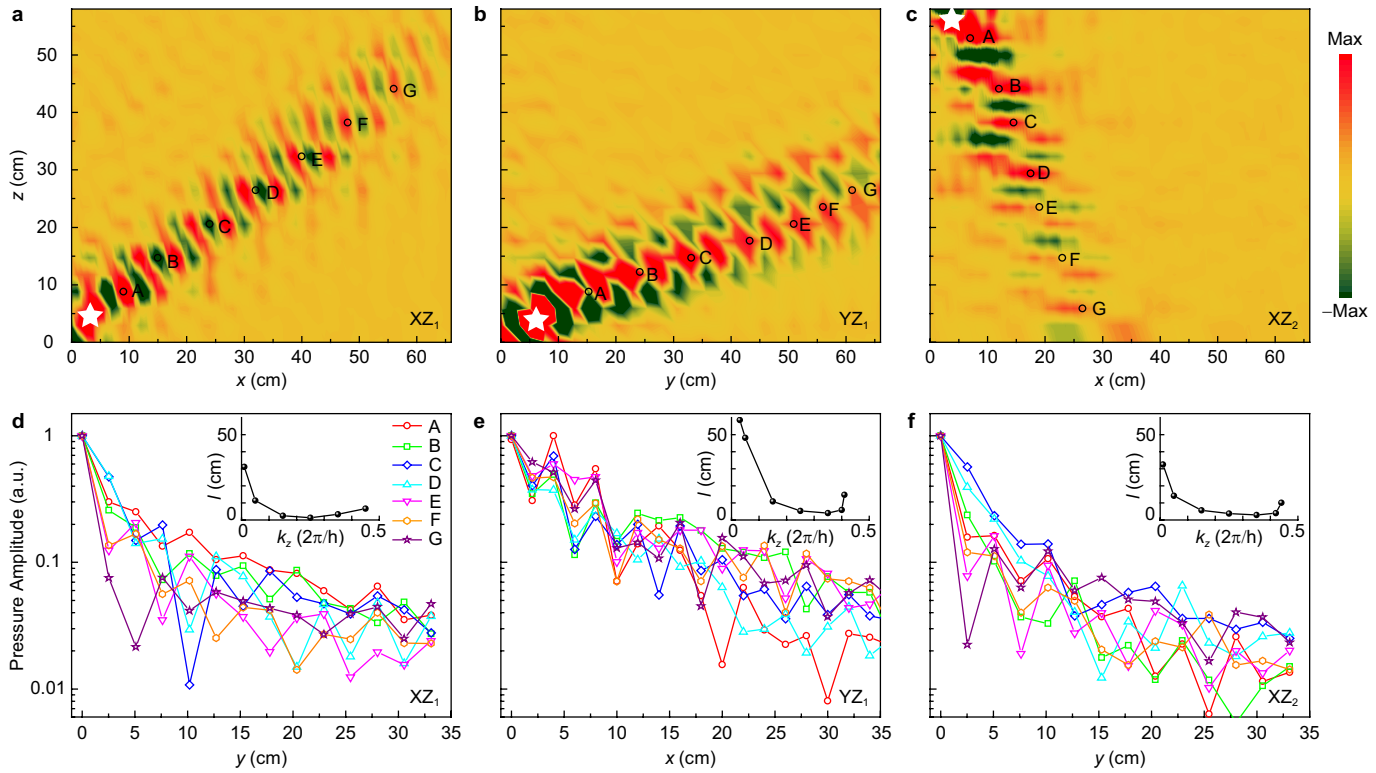
Extended Data Fig. 2 | Berry flux distributions, Chern numbers and bulk boundary correspondence. a, Berry flux contributions to the lowest two bulk bands, derived from the numerically calculated Weyl charge distributions (see Fig. 2h). Here $|C|$ labels the amplitude of the Weyl charge, and the red and black spheres indicate sources and sinks of Berry fluxes, respectively. **b**, Chern numbers for the lowest two bulk gaps opened at $k_z > 0$, labelled according to the Berry flux distributions in **a**. (Here, the band structure is simulated at $k_z = 0.5\pi/h$.) The Chern numbers C_{gap} of both gaps are -1, consistent with the topologically non-trivial surface spectra presented in Extended Data Fig. 1.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | Different EFC properties attained by engineering the surface terminations of the Weyl phononic crystal. **a**, Surface EFCs evaluated for the XZ surfaces at the Weyl frequency of 5.75 kHz, where the parameter d_y defined at the top of the column characterizes the surface truncation (black dashed line). The grey regions display the projections of bulk bands, the blue spheres display the projections of the Weyl points K and K′ at 5.75 kHz, and the arrows indicate the directions of the group velocities for the surface arc states. **b**, Similar to **a**, but for the YZ surfaces specified by d_x. The evolution of the EFCs for different surface terminations shows various possibilities of manipulating the surface states according to their group velocities. We focus on the cases in **a2**, **a4** and **b4** in the main text to attain the desired SAW properties. Throughout, we use 5.75 kHz to shrink the momentum regions projected by the bulk bands, which is favourable for the experimental observation of the surface arc states and the associated interfacial phenomena.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Experimental characterizations of the topologically protected SAWs. a–c, Pressure distributions on the sample surfaces XZ_1, YZ_1 and XZ_2 (Fig. 2d–f), respectively, scanned step by step at 5.75 kHz. The white stars indicate the positions of the point-like sound source. The propagation directions of the beams, as predicted from most of the SAWs hosted on the corresponding facets, are closely related to the positive and negative refractions observed in Fig. 3. The data are used to obtain the EFCs (with $k_z > 0$) in Fig. 2l–n through a Fourier transform. Similar data can be collected to obtain the frequency-dependent surface band dispersions at a given k_z (Fig. 2i–k). **d–f**, Decay signatures identified for the surface states in **a–c**. The data are measured along the normal directions of the sample surfaces, the in-plane coordinates of which are marked by the black circles (A–G) in **a–c**. The pressure magnitudes do not exhibit precise (oscillatory) exponential decay, mainly because the surface beam consists of many surface arc states with different out-of-plane decay lengths. The inset shows the k_z dependence of the decay length l simulated for each surface.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0339-0

Photoswitching topology in polymer networks with metal–organic cages as crosslinks

Yuwei Gu[1], Eric A. Alt[1], Heng Wang[2], Xiaopeng Li[2], Adam P. Willard[1] & Jeremiah A. Johnson[1]*

Polymer networks can have a range of desirable properties such as mechanical strength, wide compositional diversity between different materials, permanent porosity, convenient processability and broad solvent compatibility[1,2]. Designing polymer networks from the bottom up with new structural motifs and chemical compositions can be used to impart dynamic features such as malleability or self-healing, or to allow the material to respond to environmental stimuli[3–8]. However, many existing systems exhibit only one operational state that is defined by the material's composition and topology[3–6]; or their responsiveness may be irreversible[7,9,10] and limited to a single network property[11,12] (such as stiffness). Here we use cooperative self-assembly as a design principle to prepare a material that can be switched between two topological states. By using networks of polymer-linked metal–organic cages in which the cages change shape and size on irradiation, we can reversibly switch the network topology with ultraviolet or green light. This photoswitching produces coherent changes in several network properties at once, including branch functionality, junction fluctuations, defect tolerance, shear modulus, stress-relaxation behaviour and self-healing. Topology-switching materials could prove useful in fields such as soft robotics and photo-actuators as well as providing model systems for fundamental polymer physics studies.

The topological connections between polymer network strands and junctions—including branch functionality, cyclic defects, entanglements and density fluctuations—have long been implicated in dictating the bulk properties of polymer networks[13]. With the advent of new synthetic methods and tools for polymer network characterization[14–16], recent studies have revealed that network topology, as one of the few global parameters for polymer networks, plays a critical role in determining material properties such as elasticity[17], the gel point[18], network dynamics[15] and degree of defect tolerance[5]. Here we propose and examine topology switching as a strategy to develop dynamic and stimuli-responsive materials in which multiple aspects of the materials' properties are regulated coherently.

We hypothesized that polymer metal–organic cage (polyMOC) networks[5,19], which are a class of materials formed from the metallosupramolecular assembly of discrete MOCs connected to polymer chains, could provide a route to topology-switchable materials (Fig. 1a). In polyMOCs, the network junctions are nanoscale metal$_x$ligand$_y$ (M_xL_y) cages of defined shape and stoichiometry[20,21]; the MOC structure defines the polyMOC topology and bulk properties. Inspired by recent developments in modulating MOC conformation in a stimuli-responsive fashion[22,23], we began our investigation with the design and synthesis of a poly(ethylene glycol)-based polymer ligand (Fig. 1b), which features two bis-pyridyl dithienylethene (DTE) groups. We hypothesized that, in the presence of Pd^{2+}, these ligands would form MOCs that could be reversibly photoswitched between small Pd_3L_6 rings and large $Pd_{24}L_{48}$ rhombicuboctahedra by irradiating with green and ultraviolet (UV) light, respectively[23]. Before embarking on polyMOC synthesis using polymer ligands, we conducted model studies with a non-polymeric bis-pyridyl DTE-derivative ligand to confirm the photoswitching capability of these DTE derivatives (Extended Data Fig. 1).

When 1 equivalent of the open-form polymer ligand (o-PL) was mixed with 1 equivalent of $Pd(CH_3CN)_4(BF_4)_2$ in CH_3CN (25.1 mM Pd^{2+}), a dark-brown gel formed (named o-gel). After annealing at 60 °C for 4 h to enable equilibration of the network junctions, the o-gel was characterized with oscillatory rheometry (Fig. 2a, Extended Data Fig. 2a). The material displayed elastic behaviour at all frequencies with a storage modulus (G') of about 8.3 kPa at 10 rad s^{-1}. Irradiation of the o-gel with UV light for 5 h at 60 °C produced a dark blue material, which we call c-gel (for closed-form), with a nearly doubled G' (Fig. 2a). The shear loss modulus (G'') of the c-gel (Extended Data Fig. 2b) was slightly higher than that of the o-gel at the measured frequencies (1–100 rad s^{-1}). We attribute such behaviour to the presence of a greater number of topological defects (for example cyclic defects) in the c-gel—a characteristic feature for polyMOC gels with high branch functionality, f (ref. [5]). A sample of the c-gel prepared directly from pre-synthesized closed-form polymer ligand (c-PL) and Pd^{2+} displayed nearly identical frequency behaviour (Extended Data Fig. 2d), which suggests that the observed changes in mechanical properties on conversion of the o-gel to c-gel are due to photoswitching of the DTE-containing MOC junctions. Finally, exposure of the c-gel to green light for 5 h at 60 °C regenerated the o-gel with only a slight decrease in stiffness compared with the original o-gel ($G' = 7.8$ kPa; Fig. 2a, Extended Data Fig. 2c). Notably, the o-gel and c-gel have the same molar concentration of polymer strands and Pd atoms. Although the magnitude of the change in G' demonstrated here is relatively small (a factor of 1.7), in conventional supramolecular materials based on point interactions, the o-gel and c-gel would be expected to have the same G'; the hierarchical nature of polyMOCs enables access to two distinct topological states with different densities of elastically effective strands, cyclic defects and junction fluctuations, which leads to different G' .

Small-angle X-ray scattering (SAXS) was used to confirm that the mechanical changes observed above were due to changes in the MOC structure within the o-gel and c-gel (Fig. 2b, c). The scattering curve for the o-gel shows a broad peak at scattering vector $q = 0.080$ $Å^{-1}$ and a weak peak at $q = 0.729$ $Å^{-1}$ (Fig. 2b, inset). The latter peak agrees with the form factor of a uniform spherical particle with radius of 0.79 nm. We attribute this peak to scattering from Pd_3L_6 MOCs within the o-gel. The low-q SAXS peak, which corresponds to a d-spacing of about 7.85 nm, was assigned as the average distance between MOCs linked by polymer chains.

The SAXS curve for the c-gel displayed a sharpened peak shifted to lower q than for the o-gel, as well as five new peaks in the high-q regime (Fig. 2b, Extended Data Fig. 3a). The latter peaks were best fitted by the form factor of a uniform spherical particle with a 2.9-nm radius (Extended Data Fig. 3b), which agrees remarkably well with the analogous measurement by diffusion ordered NMR spectroscopy (DOSY) (3.1 nm; Extended Data Fig. 1c). The low-q SAXS peak for the c-gel corresponds to a d-spacing of 11.9 nm; this increased distance between MOCs compared with the o-gel reflects the fact that the larger

[1]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Chemistry, University of South Florida, Tampa, FL, USA. *e-mail: jaj2109@mit.edu

© 2018 Springer Nature Limited. All rights reserved.

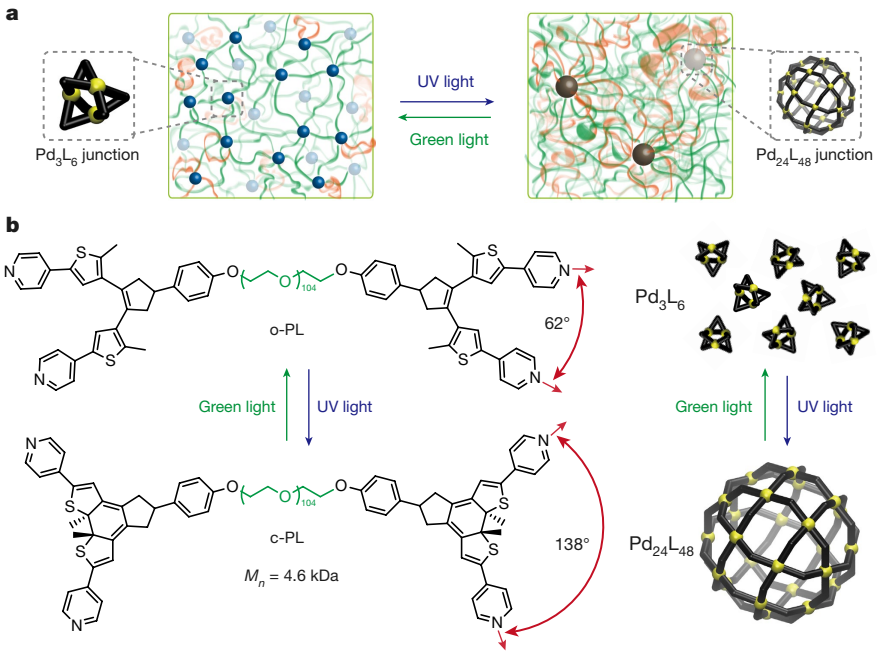


Fig. 1 | Design of polyMOCs with photoswitchable topology. **a**, Schematic illustration of photo-regulated interconversion between two different network topologies. Photoresponsive MOCs are introduced as junctions within polyMOCs. On UV irradiation, the MOC rearranges its structure from Pd_3L_6 to $Pd_{24}L_{48}$. Consequently, the network topology (for example, average branch functionality and cyclic defects) is altered. Reversal of the MOC structure with green light regenerates the original network topology. **b**, Chemical structure of the photoresponsive polymer ligand and a schematic of MOC interconversion. The DTE moiety undergoes electrocyclic ring-closure and ring-opening on UV and green-light irradiation, respectively. Note that the ring-closure reaction produces a racemic mixture of *trans*-ring-closed products. Photoinduced ring-closure/opening leads to a change of bite angle between the two attached pyridine groups. Hence, the open-form (o-PL) and closed-form (c-PL) polymer ligands form small Pd_3(o-PL)$_6$ and large Pd_{24}(c-PL)$_{48}$ MOCs, respectively, in the presence of Pd^{2+}. The existence of many diastereomeric cage structures does not seem to inhibit the assembly process.

MOCs in the c-gel must be further apart to maintain a constant Pd concentration. The sharpening and increased intensity of the low-q SAXS peak for the c-gel is another sign that this material possesses uniform Pd-rich regions ($Pd_{24}L_{48}$ MOCs).

We carried out simulations of Pd_3L_6 (o-gel) and $Pd_{24}L_{48}$ (c-gel) polyMOC gels (25.1 mM Pd^{2+}) using a coarse-grained model of dynamic network topology to probe the topology of our materials in more detail (Fig. 2d, e, Extended Data Figs. 5, 6). Specifically, the model was used to compute the density of loop defects and the average network branch functionality ($\overline{f}$).

Whereas 40.3% of the strands in the simulated o-gel were primary loops (that is, both ends of the strand were connected to the same Pd_3L_6 MOC; red strands in Fig. 2d), a remarkable 74.2% of the strands in the simulated c-gel were primary loops (red strands in Fig. 2e). Despite this increase in elastically inactive primary loops, as noted above, G' for the c-gel was 1.7 times that for the o-gel. In other words, on switching topology, the degree to which these materials could tolerate defects is changed (the c-gel can withstand more defects than the o-gel). Aiming to increase $\Delta G'$ further, we prepared the o-gel′ by replacing 15% of o-PL in the o-gel with free ligand o-L (these free ligands represent elastically inactive 'defects'), as shown in Fig. 2f. Because of its low degree of defect tolerance, the o-gel′ has much lower G' than the unmodified o-gel (about 1.2 kPa versus about 8.3 kPa). On irradiating with UV light, topology switching (see Extended Data Fig. 4d for SAXS data) results in a highly defect-tolerant state (c-gel′), increasing G' by an order of magnitude (from about 1.2 kPa to about 9.8 kPa; Fig. 2f, Extended Data Fig. 4a, b). The switching process was confirmed to be reversible (Fig. 2f, Extended Data Fig. 4c). Given such a large and reversible $\Delta G'$, we anticipate that topology-switching materials could find applications in fields such as soft robotics and photo-actuators. Moreover, the incorporation of free ligand offers a general yet simple strategy to introduce functional moieties into polyMOC junctions. As branch functionality and junction size are altered considerably on switching topology, the possible number of functional moieties that can be confined in a MOC is increased or decreased[24]; this could prove useful for reversible encapsulation and release of small molecules, as well as for responsive materials based on multivalent interactions[25].

The simulated $\overline{f}$ were compared with experimental results. To calculate experimental $\overline{f}$ from measured G' values, two network models—an affine network and a phantom network—were applied, which assume different degrees of network junction fluctuation. As the network topology switched from the o-gel to the c-gel, the local mass density around each network junction was greatly increased. Consequently, our results (Extended Data Fig. 6) indicate that the ability of the network junctions to fluctuate around average positions was diminshed, which corresponds to a reversible phantom-to-affine transition in polymer networks. Because the difference between a phantom and an affine network forms the basis of a thorough understanding of rubber elasticity[2], we anticipate that the reversible phantom-to-affine transition achieved through topology switching may be an ideal model system in which to examine these two network models experimentally.

Owing to cooperative effects, the metal–ligand bonds in large MOCs are often much less dynamic than the same bonds in smaller complexes[26]. Thus, we expect the o-gel to be much more dynamic than the c-gel, which provides the ability to switch network dynamics through topology switching. Measurements of swelling in excess CH_3CN solvent (Fig. 3a) support our hypothesis: the o-gel absorbed 102 ± 13 times its own weight in CH_3CN after 5 days and completely dissolved after 20 days, whereas the c-gel absorbed 14 ± 6 times its own weight in CH_3CN after 5 days and did not swell further over 20 days.

Stress-relaxation measurements were conducted to quantitatively evaluate the dynamics of the o-gel and c-gel as a function of temperature (Fig. 3b, c, respectively); characteristic relaxation times τ were obtained by fitting experimental data to the Kohlrausch model. For both the o-gel and c-gel, τ decreased as temperature increased, owing to thermally accelerated ligand exchange. More importantly, at the same temperature, τ for the c-gel was consistently about an order of magnitude larger than for the o-gel. In addition, τ for the o-gel at

© 2018 Springer Nature Limited. All rights reserved.

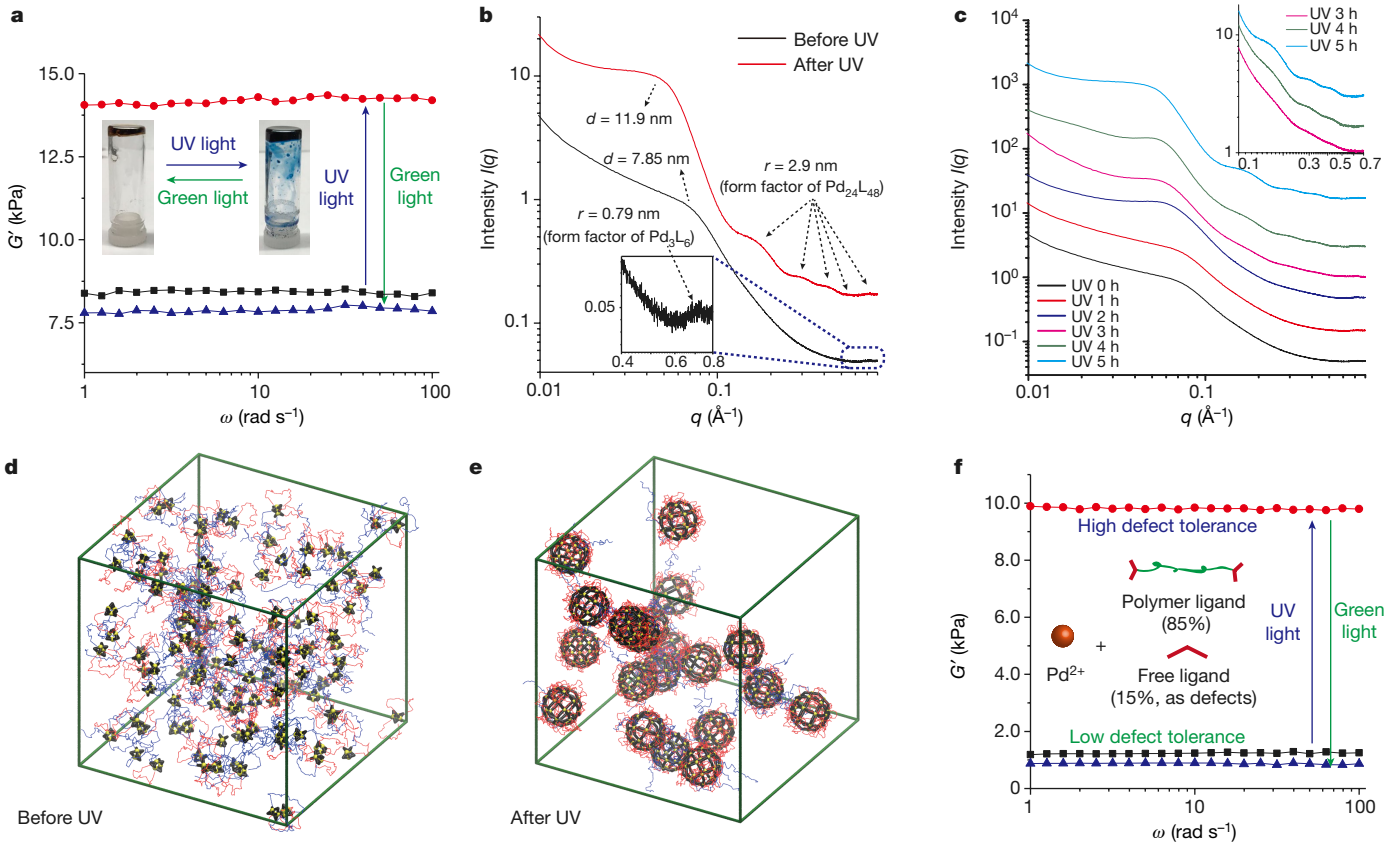


Fig. 2 | Photoswitching of polyMOC topology. a, A brown polyMOC (o-gel) was formed by mixing o-PL and $Pd(CH_3CN)_4(BF_4)_2$ in CH_3CN (25.1 mM). The shear storage modulus (G') of the o-gel was 8.3 kPa (black curve). UV irradiation of the o-gel at 60 °C for 5 h produced a blue material (c-gel) with $G' = 14.1$ kPa (red curve). Irradiation with green light regenerated the o-gel with nearly the same G' (blue curve). The slight decrease in G' on regeneration of the o-gel is believed to be due to a small amount of ligand photodegradation. **b**, SAXS curves for polyMOCs before (black curve) and after (red curve) UV irradiation. The d-spacing in the low-q regime is assigned as the average distance between MOC junctions. The high-q peaks correspond to form factors of Pd_3(o-PL)$_6$ (inset) and Pd_{24}(c-PL)$_{48}$. **c**, SAXS curves over the course of UV irradiation. Inset, zoom-in of the high-q regime of the SAXS curves. The form factor corresponding to Pd_{24}(c-PL)$_{48}$ emerges over time, providing direct evidence for the transition of network topology. **d**, Snapshot of the simulated o-gel. Primary loop strands are shown in red. **e**, Snapshot of the simulated c-gel. Primary loop strands are shown in red. **f**, Topology switching produces a marked switch in the degree of defect tolerance. By replacing 15% o-PL with o-L, we were able to switch G' by one order of magnitude, from about 1 kPa to about 10 kPa, on switching the topology. Each experimental curve in **a**, **b**, **c** and **f** shows a representative result of at least three measurements, which showed excellent reproducibility with a relative standard deviation of less than 5%.

room temperature ($\tau = 555$ s) was less than that for the c-gel at 60 °C ($\tau = 1{,}085$ s). Thanks to topology switching, the static properties (G') and dynamic properties (τ) of this material could be simultaneously switched, both by one order of magnitude. This design principle opens many potential applications, such as soft adhesives, or biomaterials for adaptive cell culture that can reversibly switch stiffness and stress relaxation rate, both of which have been shown to be essential to regulate stem-cell fate and cell proliferation[27,28].

These differences in relaxation kinetics between the o-gel and c-gel should translate to unique topology-dependent self-healing behaviours; the slow exchange kinetics of the c-gel should make it less able to heal. To test this hypothesis, samples of the o-gel and c-gel were cut into two pieces and placed in contact; considering the difference in measured relaxation timescales between the o-gel and c-gel, the self-healing condition was chosen as 4 h at 40 °C. After 4 h, the o-gel was completely healed with no visible damage at the interface, whereas the c-gel remained in two separate pieces. In addition, our polyMOC could be switched from the healable state (o-gel) to the non-healable state (c-gel) and back (Fig. 3d).

Most existing self-healing materials rely on fast bond exchange in the healable state. Although such fast bond exchange provides self-healing properties, the same mechanism leads to low mechanical robustness and considerable material deformation on loading, which makes self-healing materials unsuitable for many applications. Topology switching is a general way to break this limitation. It allows a single material to be operated in one topological state (c-gel) that is mechanically robust and suitable for heavier-duty applications; if the material is damaged or needs to be re-processed, one can simply switch its topology to the self-healable state (o-gel) to heal or process the material, and then switch back to the operation state[29].

To study the fatigue behaviour of our materials, we measured G' over several cycles of switching between the o-gel and c-gel using two sets of irradiation conditions. As shown in Fig. 4a (black line), when the polyMOC was cycled between exposure to UV light for 5 h at 60 °C and green light for 5 h at 60 °C, the material displayed elastic behaviour in both topological states for seven cycles. When cycled at lower temperature (45 °C) and longer irradiation time (12 h), the material lost its mechanical integrity after three cycles (Fig. 4a, red curve). These data are indicative of photoinduced degradation, probably of the DTE-containing ligands, because related DTE derivatives are known to degrade under extended UV irradiation[30]. SAXS was used to probe the fatigue mechanism of our polyMOCs in more detail. The scattering profiles for the o-gel and c-gel did not noticeably change after three switching cycles (Fig. 4b); however, the scattering profiles after seven switching cycles (green and yellow curves in Fig. 4b) no longer possess the MOC form factors, suggesting that UV-induced DTE degradation disrupted uniform MOC formation. The presence of low-q SAXS peaks after seven cycles suggests the presence of metal–ligand clusters

© 2018 Springer Nature Limited. All rights reserved.

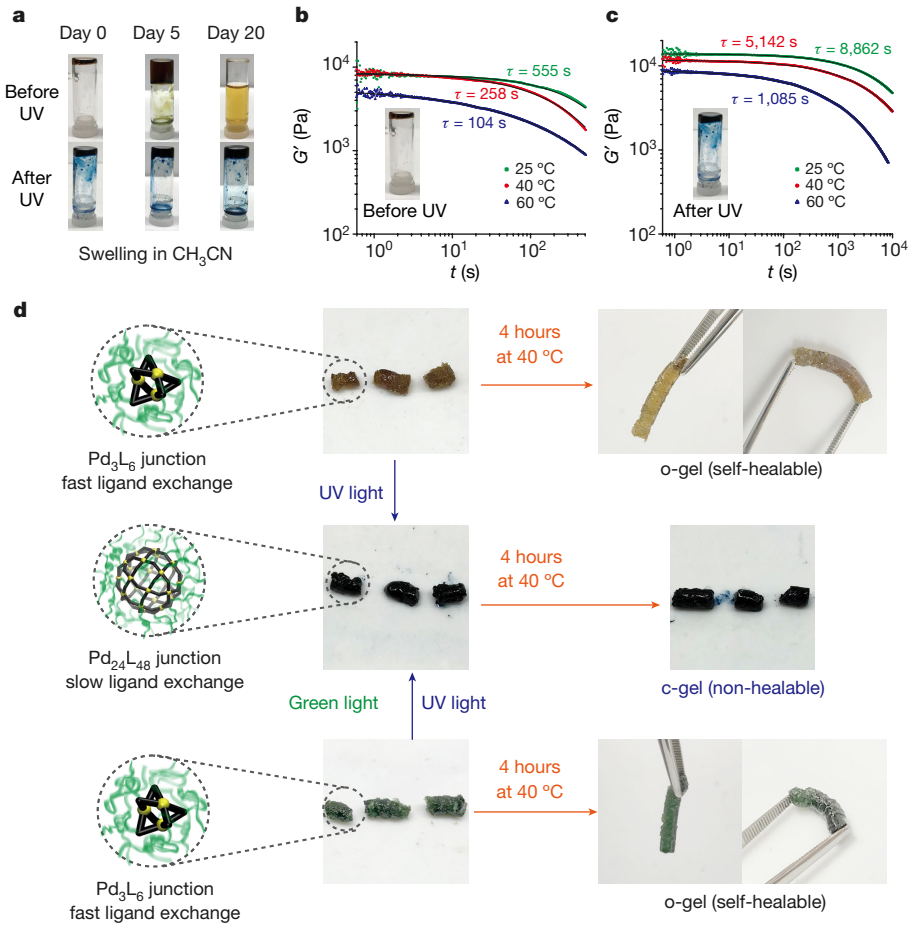


Fig. 3 | Photoswitching topology leads to tunable network dynamics. **a**, Swelling before (o-gel) and after (c-gel) UV irradiation. **b**, Stress-relaxation curves at various temperatures for the o-gel. Curves were fitted with the Kohlrausch model to provide the characteristic relaxation times τ. Symbols, raw data; solid lines, fitted data. **c**, Stress-relaxation curves at various temperatures for the c-gel. Each experimental curve in **b** and **c** shows a representative result of at least three measurements, which showed excellent reproducibility with a relative standard deviation of less than 5%. **d**, Photographs of self-healing experiments, which demonstrate that the o-gel undergoes self-healing whereas the c-gel does not.

of ill-defined structure; these clusters may still provide mechanical robustness to the material.

To elucidate the fatigue mechanism in more detail, $\overline{f}$ values were calculated based on G' values measured over several switching cycles (Fig. 4c). After seven cycles, the c-gel still possesses a very high $\overline{f}$ value (6.26 ± 0.51). In contrast, $\overline{f}$ for the o-gel after seven cycles (2.22 ± 0.07) is close to the $f = 2$ limit at which the sol–gel transition starts to occur. To model the extent of UV-induced DTE degradation, dynamic

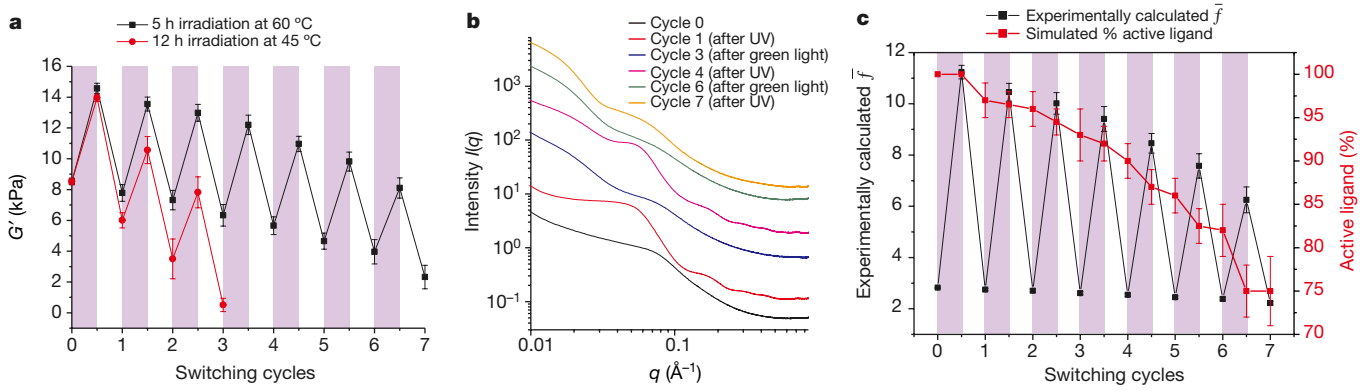


Fig. 4 | Fatigue properties. **a**, Shear modulus G' over multiple topology switching cycles. Each switching cycle includes UV irradiation (purple regions) followed by green-light irradiation (white regions). Two different irradiation schemes were applied for each UV or green-light irradiation. Higher temperature enabled reduction of the irradiation time, which gave improved fatigue resistance. Error bars represent the standard deviation of three measurements. **b**, SAXS curves for topology switching cycles. Each experimental curve shows a representative result of at least three measurements, which showed excellent reproducibility with a relative standard deviation of less than 5%. **c**, Average branch functionality ($\overline{f}$) calculated on the basis of G' over multiple switching cycles (black curves), and simulation results based on the experimentally calculated $\overline{f}$ for the ligand degradation profile over multiple switching cycles (red curves). In the calculation of $\overline{f}$, the c-gel and o-gel are treated as affine and phantom networks, respectively. Error bars represent the standard deviation of three measurements.

© 2018 Springer Nature Limited. All rights reserved.

network topology simulations were used to compute $\overline{f}$ as a function of the percentage of active ligand within these polyMOCs (Extended Data Fig. 7); $\overline{f}$ values obtained from experiment were then translated to a ligand degradation profile (Fig. 4c). From these data, we estimate that 25% of the DTE-based ligand was degraded over seven cycles. This result is consistent with the degradation profile of free ligand obtained by ^{1}H NMR (Extended Data Fig. 8).

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0339-0.

Received: 4 February 2018; Accepted: 16 May 2018;
Published online 18 July 2018.

1. Osada, Y. & Gong, J.-P. Soft and wet materials: polymer gels. *Adv. Mater.* **10**, 827–837 (1998).
2. Rubinstein, M. & Colby, R. H. *Polymer Physics* (Oxford Univ. Press, Oxford, 2003).
3. Chen, X. et al. A thermally re-mendable cross-linked polymeric material. *Science* **295**, 1698–1702 (2002).
4. Burnworth, M. et al. Optically healable supramolecular polymers. *Nature* **472**, 334–337 (2011).
5. Zhukhovitskiy, A. V. et al. Highly branched and loop-rich gels via formation of metal–organic cages linked by polymers. *Nat. Chem.* **8**, 33–41 (2016).
6. Röttger, M. et al. High-performance vitrimers from commodity thermoplastics through dioxaborolane metathesis. *Science* **356**, 62–65 (2017).
7. Cangialosi, A. et al. DNA sequence-directed shape change of photopatterned hydrogels via high-degree swelling. *Science* **357**, 1126–1130 (2017).
8. Vatankhah-Varnosfaderani, M. et al. Mimicking biological stress–strain behaviour with synthetic elastomers. *Nature* **549**, 497–501 (2017).
9. Kloxin, A. M., Kasko, A. M., Salinas, C. N. & Anseth, K. S. Photodegradable hydrogels for dynamic tuning of physical and chemical properties. *Science* **324**, 59–63 (2009).
10. Ramirez, A. L. B. et al. Mechanochemical strengthening of a synthetic polymer in response to typically destructive shear forces. *Nat. Chem.* **5**, 757–761 (2013).
11. Scott, T. F., Schneider, A. D., Cook, W. D. & Bowman, C. N. Photoinduced plasticity in cross-linked polymers. *Science* **308**, 1615–1617 (2005).
12. Rosales, A. M. & Anseth, K. S. The design of reversible hydrogels to capture extracellular matrix dynamics. *Nat. Rev. Mater.* **1**, 15012 (2016).
13. Flory, P. J. Network topology and the theory of rubber elasticity. *Br. Polym. J.* **17**, 96–102 (1985).
14. Zhou, H. et al. Counting primary loops in polymer gels. *Proc. Natl Acad. Sci. USA* **109**, 19119–19124 (2012).
15. Noda, Y., Hayashi, Y. & Ito, K. From topological gels to slide-ring materials. *J. Appl. Polym. Sci.* **131**, 40509 (2014).
16. Di Lorenzo, F. & Seiffert, S. Nanostructural heterogeneity in polymer networks and gels. *Polym. Chem.* **6**, 5515–5528 (2015).
17. Zhong, M., Wang, R., Kawamoto, K., Olsen, B. D. & Johnson, J. A. Quantifying the impact of molecular defects on polymer network elasticity. *Science* **353**, 1264–1268 (2016).
18. Wang, R., Lin, T.-S., Johnson, J. A. & Olsen, B. D. Kinetic Monte Carlo simulation for quantification of the gel point of polymer networks. *ACS Macro Lett.* **6**, 1414–1419 (2017).
19. Wang, Y. et al. Star polyMOCs with diverse structures, dynamics, and functions by three-component assembly. *Angew. Chem. Int. Ed.* **56**, 188–192 (2017).
20. Leininger, S., Olenyuk, B. & Stang, P. J. Self-assembly of discrete cyclic nanostructures mediated by transition metals. *Chem. Rev.* **100**, 853–908 (2000).
21. Harris, K., Fujita, D. & Fujita, M. Giant hollow M_nL_{2n} spherical complexes: structure, functionalisation and applications. *Chem. Commun.* **49**, 6703–6712 (2013).
22. McConnell, A. J., Wood, C. S., Neelakandan, P. P. & Nitschke, J. R. Stimuli-responsive metal–ligand assemblies. *Chem. Rev.* **115**, 7729–7793 (2015).
23. Han, M. et al. Light-controlled interconversion between a self-assembled triangle and a rhombicuboctahedral sphere. *Angew. Chem. Int. Ed.* **55**, 445–449 (2016).
24. Ueda, Y., Ito, H., Fujita, D. & Fujita, M. Permeable self-assembled molecular containers for catalyst isolation enabling two-step cascade reactions. *J. Am. Chem. Soc.* **139**, 6090–6093 (2017).
25. Kiessling, L. L. & Grim, J. C. Glycopolymer probes of signal transduction. *Chem. Soc. Rev.* **42**, 4476–4491 (2013).
26. Sato, S., Ishido, Y. & Fujita, M. Remarkable stabilization of $M_{12}L_{24}$ spherical frameworks through the cooperation of 48 Pd(ii)–pyridine interactions. *J. Am. Chem. Soc.* **131**, 6064–6065 (2009).
27. Murphy, W. L., McDevitt, T. C. & Engler, A. J. Materials as stem cell regulators. *Nat. Mater.* **13**, 547–557 (2014); erratum *Nat. Mater.* **13**, 756 (2014).
28. Chaudhuri, O. et al. Hydrogels with tunable stress relaxation regulate stem cell fate and activity. *Nat. Mater.* **15**, 326–334 (2016).
29. Fuhrmann, A. et al. Conditional repair by locally switching the thermal healing capability of dynamic covalent polymers with light. *Nat. Commun.* **7**, 13623 (2016).
30. Herder, M. et al. Improving the fatigue resistance of diarylethene switches. *J. Am. Chem. Soc.* **137**, 2738–2747 (2015).

Acknowledgements We thank the National Science Foundation (CHE-1629358 for J.A.J. and CHE-1506722 for X.L.) for support of this work. This work was supported in part by the MRSEC Program of the National Science Foundation under award number DMR-1419807. Y.G. thanks ExxonMobil for an ExxonMobil–MIT Energy Fellowship. A.P.W. and E.A.A. acknowledge funding from the Research Corporation for Scientific Advancement through a Cottrell Scholars Award. X-ray scattering experiments were performed at the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory (DE-AC0206CH11357). We thank G. Clever for discussions.

Reviewer information *Nature* thanks S. Sheiko and T. Sirk for their contribution to the peer review of this work.

Author contributions Y.G. and J.A.J. conceived the idea. Y.G. designed and conducted the synthesis and characterization studies. E.A.A. and A.P.W. conducted the simulations. H.W. and X.L. conducted the mass spectrometry studies. Y.G. and J.A.J. wrote the manuscript. All authors read and edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0339-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0339-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.A.J.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

A complete set of detailed synthetic procedures for polymer ligand (PL) and spectral data is available in Supplementary Information.

Materials and instrumentation. We purchased all deuterated solvents from Cambridge Isotope Laboratories; 2-chloro-5-methylthiophene from Alfa Aesar; and 4-bromopyridine hydrochloride from Ark Pharm. All other reagents and solvents were purchased from Sigma-Aldrich. Anhydrous, degassed dichloromethane (DCM) and tetrahydrofuran (THF) were used from a J.C. Meyer solvent purification system. High-performance liquid chromatography (HPLC) grade DCM and THF were sparged vigorously with argon for at least one hour before being connected to the solvent purification system. All reactions were performed using standard Schlenk techniques and anhydrous solvents.

UV irradiation was performed with 8 W SANKYO DENKI Blacklight UV lamps with peak emission at 300 nm. Visible light irradiation was performed with four 3-W Sunlite 80146 green LED light bulbs.

All chromatography was performed on EMD Millipore silica gel 60, particle size 0.040–0.063 mm (230–400 mesh), on a Biotage Isolera Prime flash purification system. Gel permeation chromatography (GPC) was performed on an Agilent 1260 LC stack equipped with an Agilent multi-wavelength UV–vis detector, a Wyatt TrEX refractive index detector, a Wyatt DAWN EOS 18-angle light scattering detector and two Shodex KD-806M GPC columns. The GPC system was set to 60 °C at a flow rate of 1 ml min^{-1} with 0.025 M LiBr in DMF. UV–vis spectra were obtained from a Varian Cary 50 Scan UV–vis spectrophotometer.

We acquired ^{1}H nuclear magnetic resonance (^{1}H NMR) and ^{13}C nuclear magnetic resonance (^{13}C NMR) spectra on a 500-MHz Varian INOVA spectrometer and processed them in MestReNova 11.0.4. We acquired ^{1}H DOSY spectra on the same spectrometer. The gradients were increased from 3.6 G cm^{-1} to 29.2 G cm^{-1} in equally spaced steps using 16 scans. Gradient pulse (δ) was set to 2.0 ms and diffusion time (Δ) set to 0.1 s. Diffusion coefficients for resolved ^{1}H signals were extracted from decay curves using the 'peak height fit' in DOSY Transform module of MestReNova 11.0.4.

High-resolution mass spectrometry was done on a Waters quadrupole time-of-flight Premier instrument. Each sample was dissolved by CH_3CN and analysed by direct flow injection (injection volume = 1 μl or 5 μl) electrospray ionization (ESI) in the positive mode. The optimized condition was as follows: ESI capillary voltage 3,000 V, sample cone voltage 35 V, source temperature 120 °C and desolvation temperature 350 °C.

Mass spectrometry of the solution self-assemblies between Pd^{2+} and ligand was recorded with a Waters Synapt G2 spectrometer, using 0.5-mg complex sample in 1 ml of CH_3CN/CH_3OH (3:1, v/v). The following parameters were used: ESI capillary voltage 3.0 kV; sample cone voltage 20 V; extraction cone voltage 0.1 V; source temperature 100 °C; desolvation temperature 100 °C; desolvation gas flow 700 l h^{-1} (N_2); sample flow rate 10 μl min^{-1}.

Frequency sweep rheology experiments were performed on a TA Instruments Discovery Hybrid Rheometer HR-2. The rheometer was outfitted with an active temperature control system with an environmental enclosure for temperature control. A parallel-plate geometry (radius 8 mm) was used and coupled with a bottom plate, with the typical gap of 1.00 mm between the two plates. Frequency sweep experiments were performed from 1 rad s^{-1} to 100 rad s^{-1} at 0.5% strain, which was first confirmed to be in the linear viscoelastic regime using strain sweep experiments. Shear modulus G' was determined based on G' values at 10 rad s^{-1}. Stress-relaxation experiments were carried out at 2% strain. The loaded gel samples were immersed in mineral oil to reduce solvent evaporation and gel de-swelling by adsorbing moisture. The stress relaxation curves were fitted to a stretched exponential function $G = G_0 e^{-(t/\tau)^{\alpha}}$, from which a characteristic relaxation τ could be obtained[19]. In all the fittings performed, R^2 was greater than 0.99.

Transmission SAXS was conducted at beamline 12-ID-B at the Advanced Photon Source at Argonne National Laboratory. The photon energy was 14 keV (that is, wavelength was 0.8857 Å), the beam size was 60 × 200 μm^2, and the detectors used were Pilatus 2M (SAXS) and Pilatus 300K (wide-angle X-ray scattering WAXS). The sample-to-detector distances were calibrated using silver behenate (AgBe); for the SAXS detector, this was 3,612.22 mm. The exposure time was set to 0.2 s during data collection. A small amount of polyMOCs was removed with a spatula to fill the hole of a circular washer that acted as a sample holder (outer diameter 24 mm, inner diameter 2 mm, thickness 1 mm). To minimize the solvent evaporation on such small-volume samples, the measurement was conducted immediately after the sample was loaded onto the washer.

Swelling measurements were carried out in CH_3CN: 0.3 ml of as-formed polyMOC (before or after UV irradiation, 25.1 mM PL concentration) in a 4-ml vial was swollen in CH_3CN at room temperature. After 5 days, the CH_3CN that was not absorbed by the polyMOC was decanted. The inner wall of the vial was carefully wiped to remove the solvent adhering to the wall. The weight of the gel was measured by subtracting the weight of the vial. The polyMOC was further swollen in CH_3CN for 15 days. The same procedure was used to measure its swollen weight at day 20. The swelling ratio was calculated as the ratio between the swollen weight and dry weight, with the latter estimated on the basis of the amount of polymer ligand and $Pd(CH_3CN)_4(BF4)_2$ added during the synthesis of the polyMOC.

Solution self-assembly of model ligand with Pd^{2+}. We used 1,2-bis(2-methyl-5-(pyridin-4-yl)thiophen-3-yl)cyclopent-1-ene (o-L) as a model compound to study the photo-controlled self-assembly of our non-fluorinated DTE bis-pyridine ligand and Pd^{2+}.

To form Pd_3(o-L)$_6$ complex, we mixed 2.8 ml o-L solution (4.29 mM in CD_3CN) with 0.2 ml $Pd(CH_3CN)_4(BF4)_2$ solution (30.0 mM in CD_3CN). On mixing, the solution quickly turned to light green. The mixture was stirred at 70 °C for one hour.

To form the Pd_{24}(c-L)$_{48}$ complex, 2.8 ml o-L solution (4.29 mM in CD_3CN) was first irradiated with UV light for one hour to form a dark purple solution. ^{1}H NMR confirmed that all the o-L was converted to c-L. We then added 0.2 ml $Pd(CH_3CN)_4(BF4)_2$ solution (30.0 mM in CD_3CN) to the aforementioned c-L solution. On mixing, the solution quickly turned to dark blue. The mixture was stirred at 70 °C for one hour.

Photoswitching from Pd_3(o-L)$_6$ to Pd_{24}(c-L)$_{48}$ was performed by irradiating Pd_3(o-L)$_6$ solution (2 mM Pd^{2+} concentration in CD_3CN) with UV light for 15 h. Photoswitching from Pd_{24}(c-L)$_{48}$ to Pd_3(o-L)$_6$ was performed by irradiating Pd_{24}(c-L)$_{48}$ solution (2 mM Pd^{2+} concentration in CD_3CN) with green LED light for 10 h. It should be noted that the conversion from Pd_3(o-L)$_6$ to Pd_{24}(c-L)$_{48}$ is slow. The same observation was made for an analogous system[23]. These authors postulated that only the small amount of free o-L present at equilibrium in a solution of Pd_3(o-L)$_6$ is able to undergo photocyclization. Under gradual but slow shifting of the dynamic equilibrium, the latter reaction delivers the c-L that subsequently reacts with Pd^{2+} to Pd_{24}(c-L)$_{48}$.

From DOSY spectra, we measured that the diffusion value (D) for Pd_3(o-L)$_6$ is $(4.65 \pm 0.51) \times 10^{-10}$ m^2 s^{-1}, and that for Pd_{24}(c-L)$_{48}$ is $(1.94 \pm 0.26) \times 10^{-10}$ m^2 s^{-1} (Extended Data Fig. 1c).

To calculate the hydrodynamic radii, we used the Stokes–Einstein equation:

$$r = \frac{kT}{6\pi\eta D}$$

where r is radius, k is Boltzmann constant, T is temperature, η is dynamic viscosity of CD_3CN (3.69×10^{-4} Pa s^{-1})[23] and D is diffusion value estimated by the DOSY experiment.

As a result, the hydrodynamic radius was estimated to be 1.27 nm for Pd_3(o-L)$_6$ and 3.05 nm for Pd_{24}(c-L)$_4$.

Synthesis of polyMOCs. *o-gel*. In a 4-ml vial, 49.0 mg o-PL (MW = 5,576) was dissolved in 210 μl CH_3CN, to which 90 μl $Pd(CH_3CN)_4(BF4)_2$ solution (97.6 mM in CH_3CN) was added. The mixture was vortexed for 20 s before it was annealed at 60 °C for 4 h to obtain the o-gel. Such gels were used for SAXS studies, swelling tests and self-healing tests. To calculate the concentration of PL (or Pd^{2+}) in the gels, the volume of poly(ethylene glycol) is taken into account (the density of poly(ethylene glycol) is taken as 1.0 g cm^{-3}), and the concentration of PL is thus 25.1 mM.

To prepare samples for shear rheology measurement, the as-mixed gels were transferred to a Teflon mould (8 mm in diameter, 2 mm in height) and sandwiched between two other Teflon sheets assisted by paper clips[19]. The assembly was then heated in an oven at 60 °C for 4 h. A clear circular gel pad was obtained and loaded on the rheometer.

c-gel. We prepared a CH_3CN solution of c-PL by UV irradiation of o-PL solution for one hour and no further purification. The c-gel was prepared analogously to the o-gel by replacing o-PL with c-PL.

o-gel′. This material was prepared in a similar fashion to the o-gel, with 15% o-PL replaced by o-L (note that each o-PL has two ligands whereas each o-L only has one ligand). The stoichiometry between Pd^{2+} and total amount of ligand was controlled at 1:2.

Photoswitching between the o-gel and c-gel. To aid the diffusion of polymer-bound ligands as well as the conversion from Pd_3(o-L)$_6$ to Pd_{24}(c-L)$_{48}$ (which strongly relies on the dynamics and equilibrium of ligand exchange), the light irradiation was performed at elevated temperature. The o-gel was placed in a closed vial and placed on its side on a digitally controlled hot plate. The surface temperature of the hot plate was set to 60 °C and closely monitored with an infrared thermometer. The conversion from the o-gel to c-gel was achieved in 5 h of UV irradiation. Similarly, when the c-gel was irradiated with green light for 5 h at 60 °C, conversion to the o-gel was achieved. We used 5 h light irradiation at 60 °C in most of this work, except for some of the gels prepared for fatigue studies, for which 12 h light irradiation at 45 °C was used.

For the samples used in shear rheology measurements, the sandwiched-assembly method was again used to prevent the deformation of gel samples. Briefly, the gels were transferred to a Teflon mould (8 mm in diameter, 2 mm in height) and sandwiched between two transparent glass slides assisted by paper clips. The assembly

© 2018 Springer Nature Limited. All rights reserved.

was then placed on a hot plate at 60 °C and was irradiated for 5 h with UV light (or green light).

SAXS characterization of topology switching. The high-q regime of the SAXS profile for the c-gel was assigned as the form factor of Pd_{24}(c-L)$_{48}$, and the peaks were thus fitted with the form factor of a solid spherical particle[31], in which R is the radius of the sphere:

$$P(q) \propto \frac{[\sin(qR) - qR\cos(qR)]^2}{(qR)^6}$$

The fitting results (Extended Data Fig. 3) show that the form factor of a spherical particle with radius 2.9 nm agrees well with the five observed peak positions.

Simulation method. Each polymer chain has two ends, and a system's configuration can be expressed as a function of the binding site positions, $\boldsymbol{r}$, and the connectivity between binding site positions via the polymers. Because each binding site is connected to at most one other binding site, the polymer topology connecting binding sites can be expressed as an adjacency vector $\boldsymbol{A}$:

$$\boldsymbol{A} = (A_1, A_2, \ldots, A_{N-1}, A_N)^\top$$

where N is the number of network node binding sites available to the polymer chain ends. If the two ends of a polymer are bound at sites i and j, then the sites are connected, and $A_i = j$ and $A_j = i$. It is also possible for one end of a polymer chain to be bound to a network node binding site at a fixed position while the other end of the chain is not bound to a network node (it is not at one of the sites $\in [1, N]$ with a defined position). In this case we define $A_i = 0$. The entropic potential for the polymer chains is the same model for polyethylene glycol described in previous work[5], and the probability for a polymer chain of length l to have an end-to-end distance r is denoted $P_l(r)$.

To sample a network configuration, nodes are generated at random positions within a simulation volume subject only to an excluded volume constraint to prevent nodes from overlapping until a desired system density is reached. Ligand binding site positions are arranged around each node position according to the geometry of the node. Starting with all of the binding positions being available, $a = \{1, \ldots, N\}$, and no connections, $\boldsymbol{A} = \boldsymbol{0}$, available positions are chosen one at a time at random, removed from the list of available positions, and paired with another available binding position (if a viable one exists), which is then also removed from the list of available binding positions. This process continues until all the binding sites at network nodes have been paired up (except in cases when no viable pair exists). The probability that the ligand on one end of a polymer chain will bind at position $\boldsymbol{r}_j$ given that the ligand on the other end is at position $\boldsymbol{r}_i$ is proportional to the probability of the free polymer of length l having an end-to-end distance r_{ij}. The probability of position i pairing with available position j (assuming there is at least one viable pair available to i) is:

$$P(A_i = j) = \frac{P_l(r_{ij})}{\sum_{k \in a} P_l(r_{ik})} \quad \text{if } \exists\, k \in a \text{ such that } P_l(r_{ik}) > 0$$

Simulations used a cubic, fully periodic simulation cell with side lengths of 100 nm, containing 16,154 pairs of connected ligands mediated by a polymer potential parameterized for polyethylene glycol with 100 repeat units. The network nodes for a given simulation comprised 4, 6, 12, 24 or 48 binding positions arranged in the Pd_2L_4, Pd_3L_6, Pd_6L_{12}, $Pd_{12}L_{24}$ or $Pd_{24}L_{48}$ geometry, respectively. Results were averaged over 100 independent simulations for each set of conditions. In all the simulations performed, the concentration of Pd^{2+} was set to 25.1 mM and the stoichiometry between Pd^{2+} and PL was set to 1:1 (that is, $[Pd^{2+}]$:[L] = 1: 2), which is consistent with the experimental condition.

Phantom-to-affine network transition through topology switching. Along with the increase in primary loops described in main text, simulation results indicate that the ultra-high functionality of the c-gel leads to an abundance of secondary loops (Extended Data Fig. 5b), wherein two MOCs are bridged by two or more strands. Unlike primary loops, which are elastically inactive, secondary loops do contribute to elasticity; however, the extent of their contribution depends on which theoretical model network is considered (for example phantom or affine network)[17,32]. To address which theory was more appropriate to describe each topological state of our material, we calculated the average total number of connections per MOC ($\overline{f}_{tc}$, an affine model-based approximation for $\overline{f}$; Extended Data Fig. 5c) and the average number of active connections (where secondary loops are treated as one effective connection; Extended Data Fig. 5d) per MOC ($\overline{f}_{ac}$, a phantom model-based approximation), as a function of MOC stoichiometry for a range of simulated polyMOCs (Extended Data Fig. 6). We found that for polyMOCs derived from small MOCs, for example the o-gel, the difference between $\overline{f}_{tc}$ and $\overline{f}_{ac}$ is almost negligible. In contrast, the large number of secondary loops in the c-gel leads to an $\overline{f}_{tc}$ value that is three times larger than $\overline{f}_{ac}$. We postulated that in the c-gel topological state, the high density of polymer chains attached to each $Pd_{24}L_{48}$ junction may suppress junction fluctuations, which would lead to affine network behaviour. Moreover, junction fluctuations in the o-gel state would be best captured by the phantom network model[5,17]. In other words, our photoswitchable polyMOC could behave as two fundamentally different networks depending on its topological state.

This hypothesis was supported by our rheological data. Measured G' values for the o-gel and c-gel were used to calculate $\overline{f}_{affine}$ and $\overline{f}_{phantom}$ (see Supplementary Information). As shown in Extended Data Fig. 6, $\overline{f}_{phantom}$ agrees well with $\overline{f}_{ac}$ for the o-gel. In contrast, $\overline{f}_{affine}$ shows excellent agreement with $\overline{f}_{tc}$ for the c-gel.

Fatigue studies. To study the fatigue properties of polyMOCs over switching cycles, we prepared several samples of the o-gel using the sandwiched-assembly method, where the circular gel in a Teflon mould was sandwiched between two transparent glass slides and annealed. Each o-gel was irradiated alternately with UV light and green light for several cycles (from zero to seven cycles) at 60 °C (each irradiation took 5 h) or 45 °C (each irradiation took 12 h). The polyMOC was then taken out of the mould and subjected to rheological or SAXS measurements. For each data point, averages and standard deviations were computed for three trials. To avoid the deformation of samples or solvent loss, samples were not re-used after measurements (that is, if seven switching cycles were examined, a total of 15×3 gel samples were made using the sandwiched-assembly method).

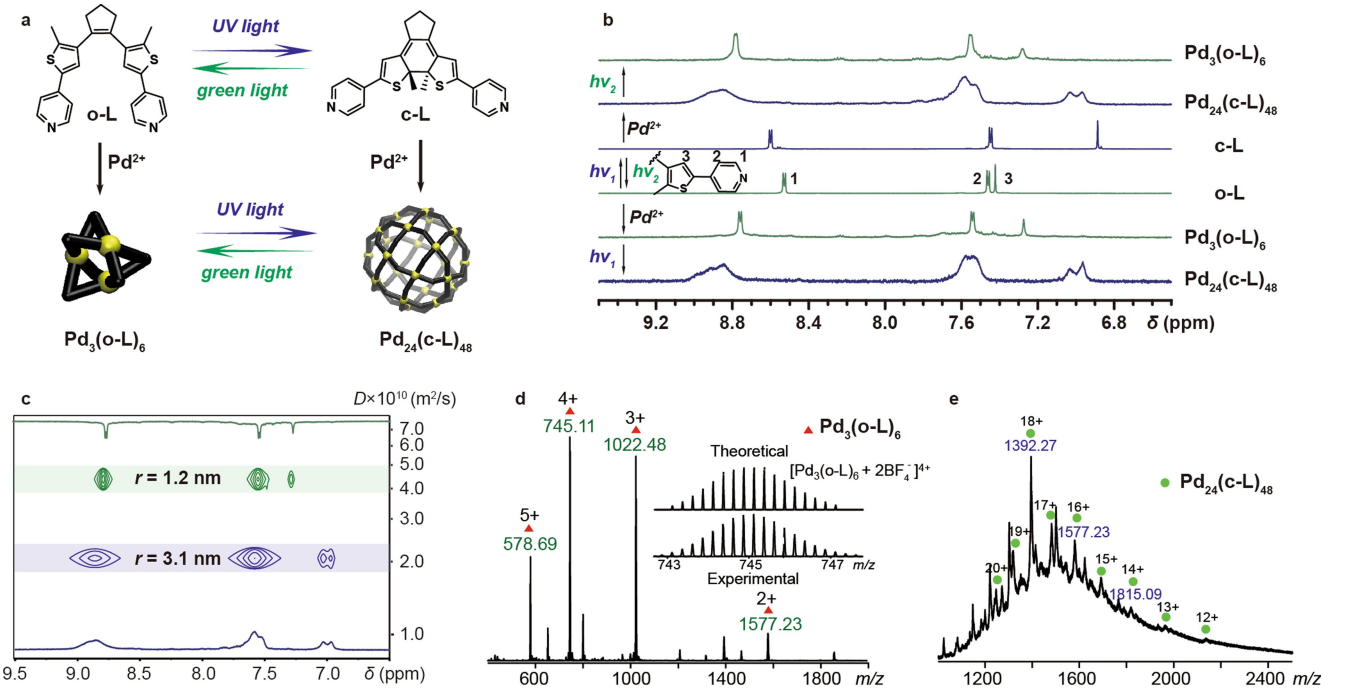
To simulate the fatigue behaviours, we set a certain fraction of the ligand as inactive as a starting parameter, then simulated the polyMOC self-assembly process and obtained the average branch functionality in the equilibrium stage. Given the postulation that the o-gel adopts phantom network behaviour while the c-gel adopts affine network behaviour, the simulated branch functionality for the o-gel is based on active connections (Extended Data Fig. 7a) and for the c-gel is based on total connections (Extended Data Fig. 7b).

From our study on the small-molecule ligand (Extended Data Fig. 8), about 80% of the ligand is still active after 18 h of UV irradiation. On the other hand, after seven switching cycles, the polyMOC has been exposed to UV irradiation for 35 h in total. We postulate that the fatigue behaviour of our polyMOCs is mainly dictated by UV-induced side reactions of DTE.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

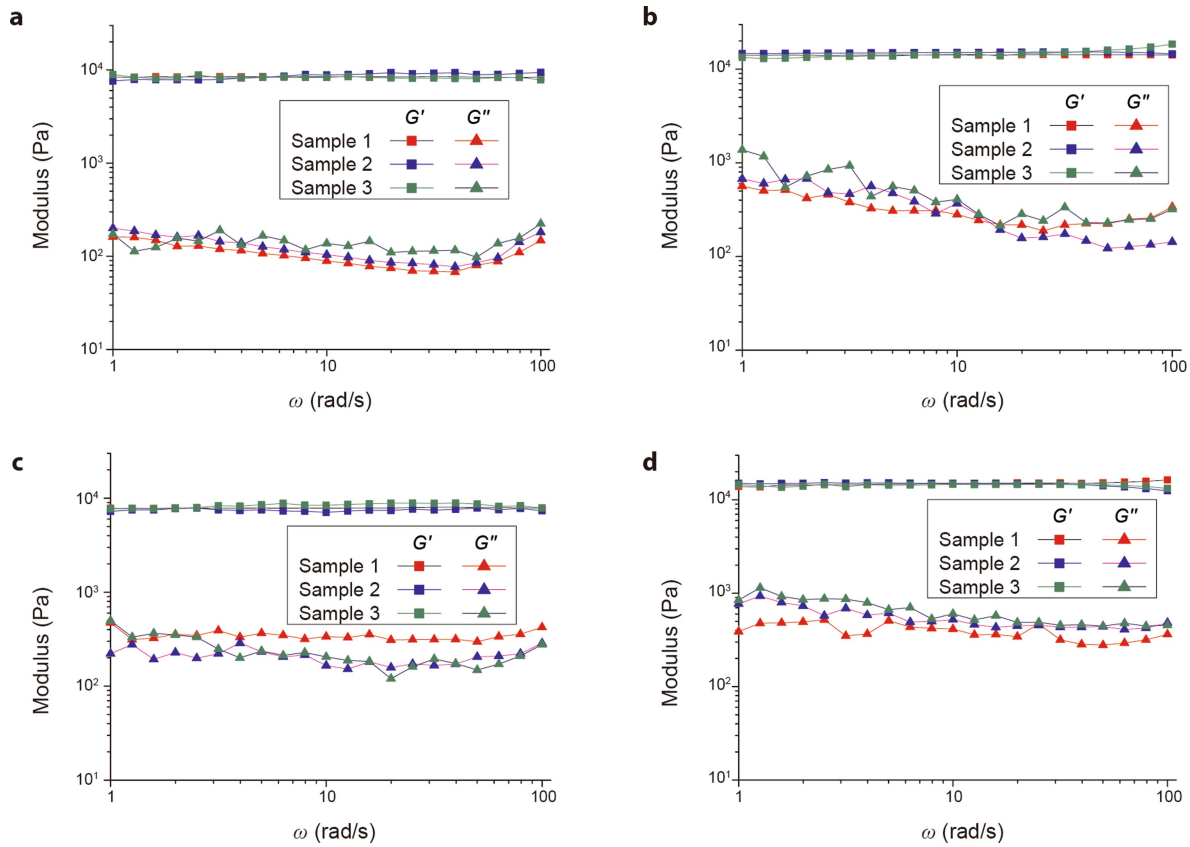
31. Pedersen, J. S. Analysis of small-angle scattering data from colloids and polymer solutions: modeling and least-squares fitting. *Adv. Colloid Interface Sci.* **70**, 171–210 (1997).
32. Lang, M. Elasticity of phantom model networks with cyclic defects. *ACS Macro Lett.* **7**, 536–539 (2018).

© 2018 Springer Nature Limited. All rights reserved.



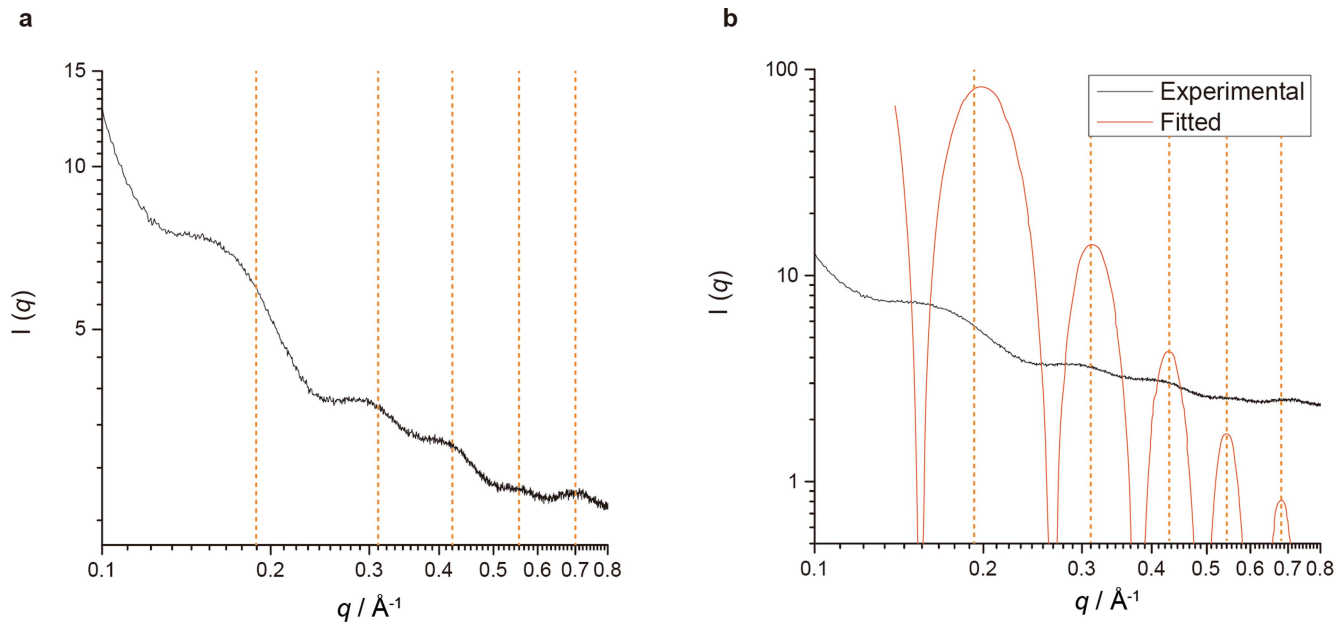
Extended Data Fig. 1 | Solution self-assembly of DTE-based bis-pyridine ligand and Pd^{2+}. **a**, DTE-containing bis-pyridine photoswitch reversibly interconverts between open (o-L) and closed (c-L) forms. Note that the ring-closure reaction produces *trans*-isomers (racemic). In the presence of Pd^{2+}, o-L and c-L form $Pd_3(o\text{-}L)_6$ and $Pd_{24}(c\text{-}L)_{48}$ MOCs, which can be interconverted using light. **b**, Aromatic regions of the solution ^1H NMR spectra (CD_3CN, 25 °C, $\omega/2\pi = 500$ MHz) for o-L and c-L, and corresponding MOCs prepared from these ligands and $Pd(CH_3CN)_4(BF_4)_2$. Photoswitching steps are indicated by black arrows. **c**, ^1H DOSY measurements indicate that Pd^{2+} forms a smaller assembly with o-L (green spectrum) and a larger assembly with c-L (blue spectrum). From a series of ^1H NMR spectra generated from ^1H DOSY measurements, the decay of peak intensities was fitted to provide the diffusion value (D) for $Pd_3(o\text{-}L)_6$ (green shaded region) and $Pd_{24}(c\text{-}L)_{48}$ (blue shaded region), respectively. From the measured diffusion values, the hydrodynamic radii of $Pd_3(o\text{-}L)_6$ and $Pd_{24}(c\text{-}L)_{48}$ were calculated. **d**, ESI mass spectrum of $Pd_3(o\text{-}L)_6$. The charge states of intact assemblies due to the loss of counterions are marked. Inset shows the simulated and observed isotopic patterns of $[Pd_3(o\text{-}L)_6 + 2BF_4^-]^{4+}$. **e**, Cold spray ionization mass spectrum of $Pd_{24}(c\text{-}L)_{48}$. The charge states of intact assemblies due to the loss of counterions are marked.

© 2018 Springer Nature Limited. All rights reserved.



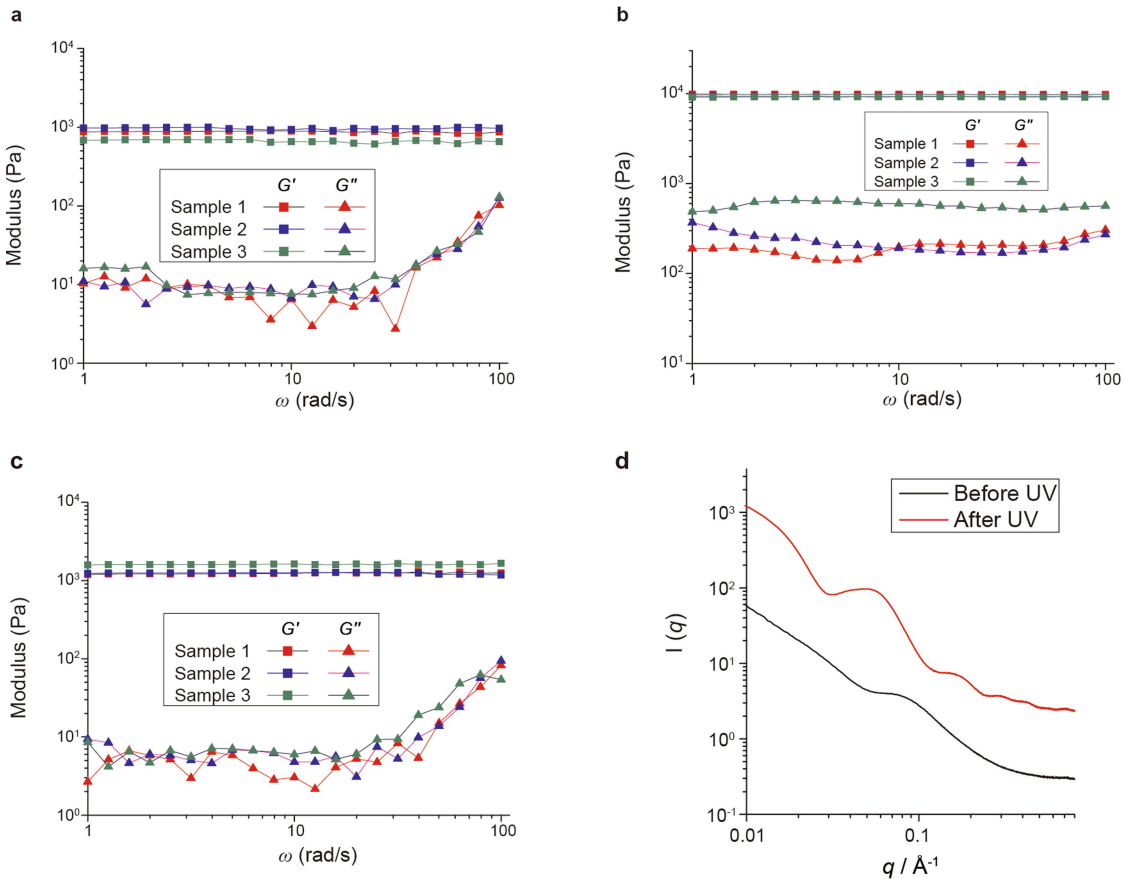
Extended Data Fig. 2 | Frequency sweeps in oscillatory rheometry at 0.5% strain. a, Data for three o-gel samples. **b**, Data for three o-gel samples after UV irradiation. **c**, Data for three o-gel samples after UV irradiation followed by green-light irradiation. **d**, Data for three c-gel samples prepared directly from c-PL.

© 2018 Springer Nature Limited. All rights reserved.



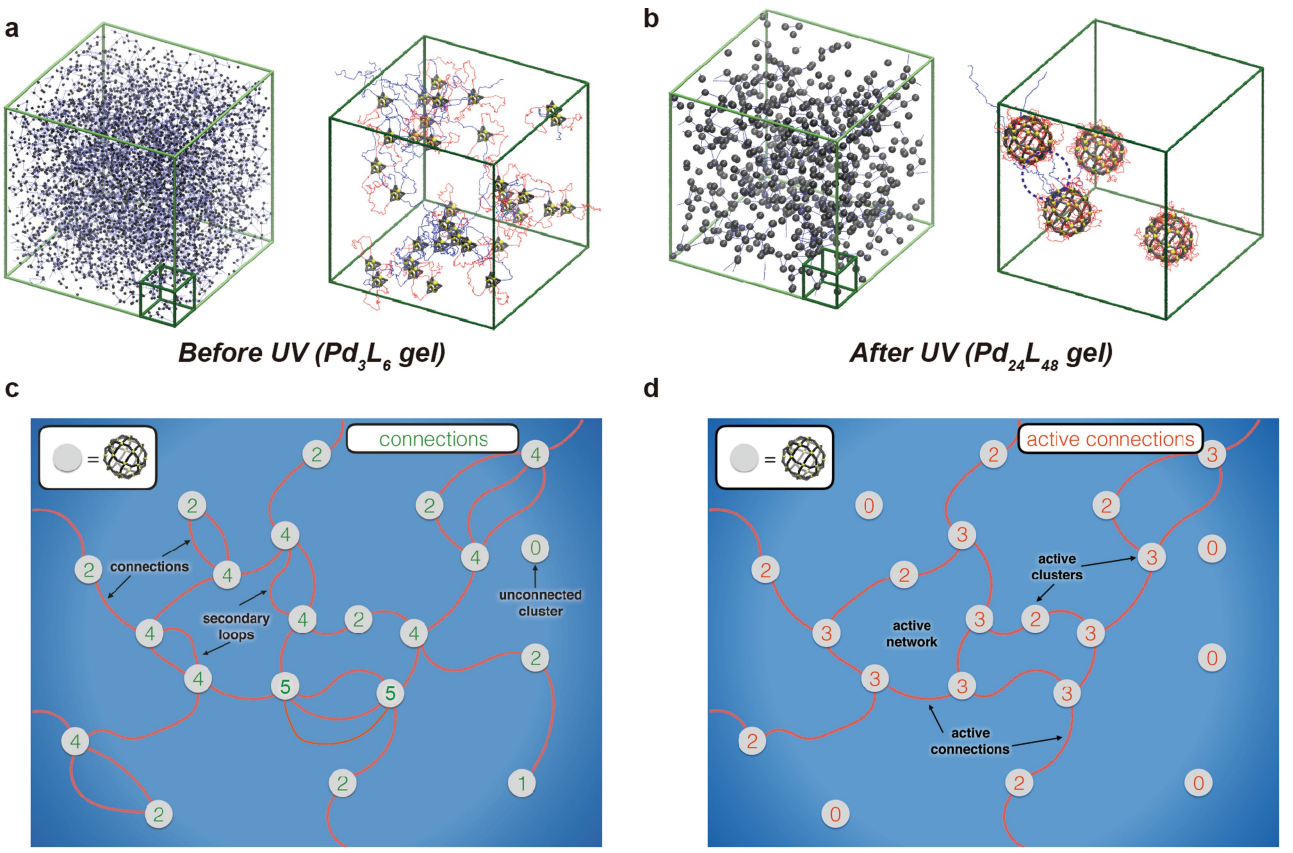
Extended Data Fig. 3 | Fitting of high-*q* SAXS profile of the c-gel. a, High-*q* regime of the SAXS profile for the c-gel. Five peaks were identified (dashed lines) and indexed. **b**, Experimental results were fitted with the form factor of a spherical particle with radius 2.9 nm.

© 2018 Springer Nature Limited. All rights reserved.



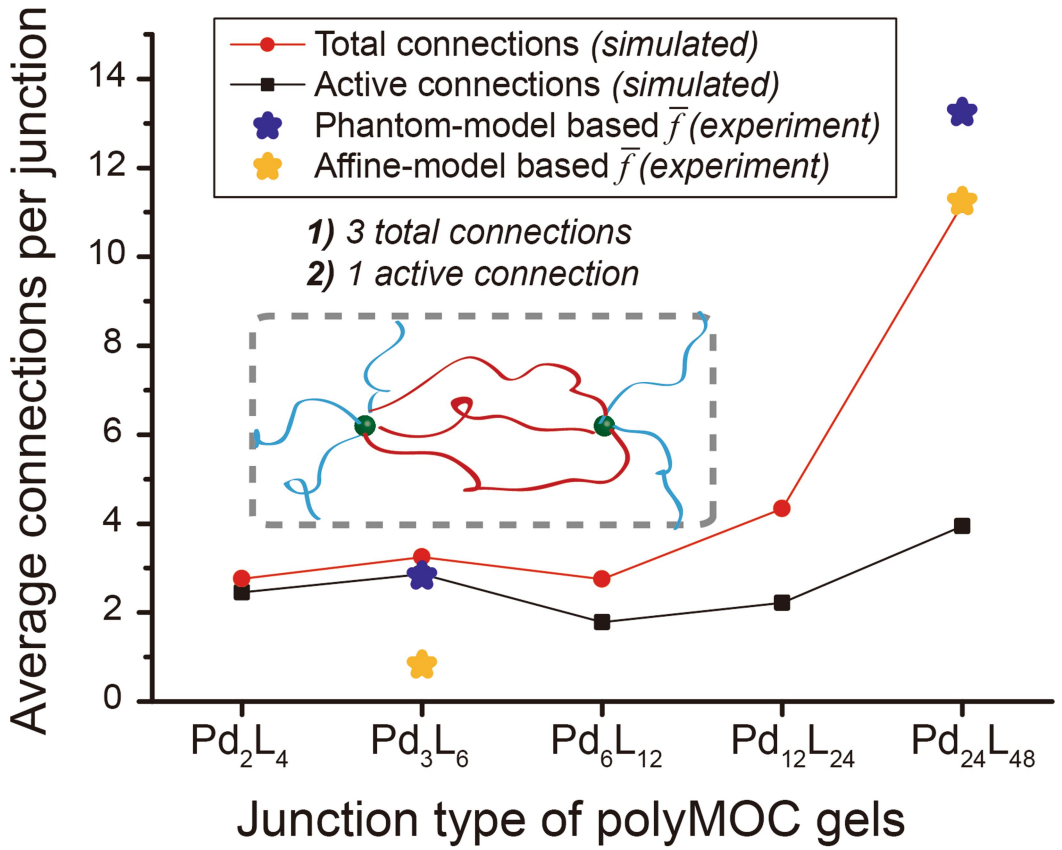
Extended Data Fig. 4 | Topology switching in the presence of free ligand as defects. a, Frequency sweep in oscillatory rheometry at 0.5% strain for three o-gel′ samples. **b**, Frequency sweep in oscillatory rheometry at 0.5% strain for three o-gel′ samples after UV irradiation. **c**, Frequency sweep in oscillatory rheometry at 0.5% strain for three o-gel′ samples after UV irradiation followed by green-light irradiation. **d**, SAXS curves for the o-gel′ before and after UV irradiation.

© 2018 Springer Nature Limited. All rights reserved.



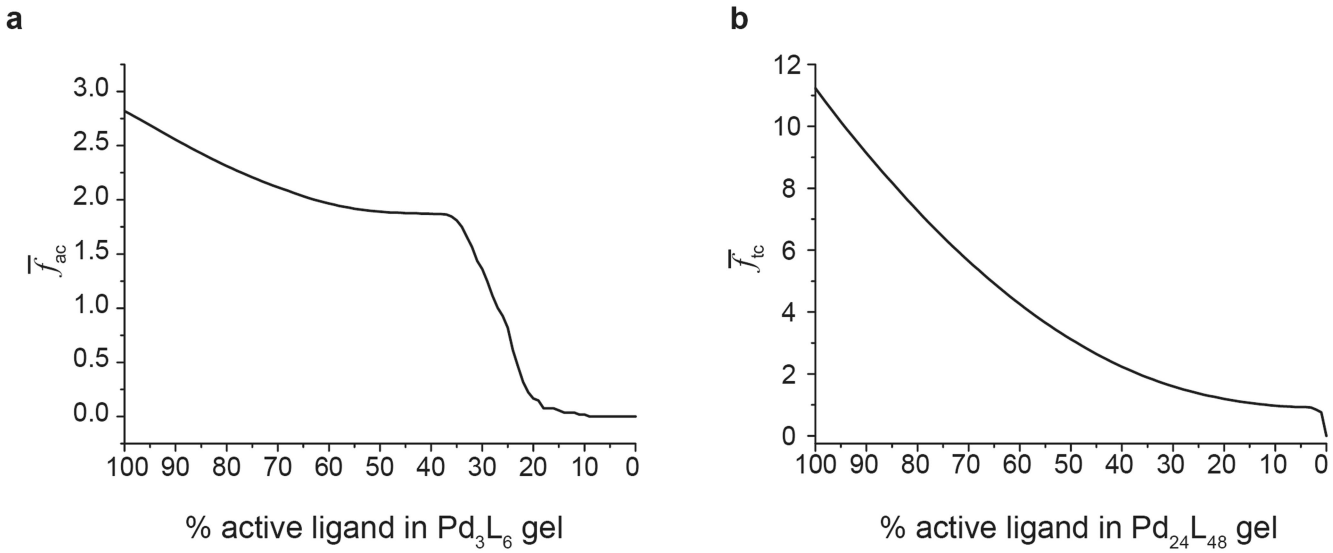
Extended Data Fig. 5 | Simulations of network topology. a, Snapshots of in silico polyMOCs before UV irradiation. Left: MOC junctions are shown as grey spheres and polymer chains connecting MOC junctions are shown in blue. Right: A zoom-in view of the region in the green cube in the left panel. Looped and non-looped polymer chains are shown in red and blue, respectively. **b**, Snapshots of in silico polyMOCs after UV irradiation. Left, MOC junctions are shown as grey spheres, and polymer chains connecting MOC junctions are shown in blue. Right, A zoom-in view of the region in the green cube in the left panel. Looped and non-looped polymer chains are shown in red and blue, respectively. The blue dashed circle shows a representative case in which two $Pd_{24}L_{48}$ clusters are connected by multiple polymer chains (that is, multiple secondary loops). **c**, A representative polymer network which is abundant in secondary loops. The connectivity of each junction is calculated on the basis of the total connections, which describes the number of polymer chains connecting MOCs. **d**, The same polymer network is represented in another way by calculating the connectivity of each junction on the basis of the active connections.

© 2018 Springer Nature Limited. All rights reserved.



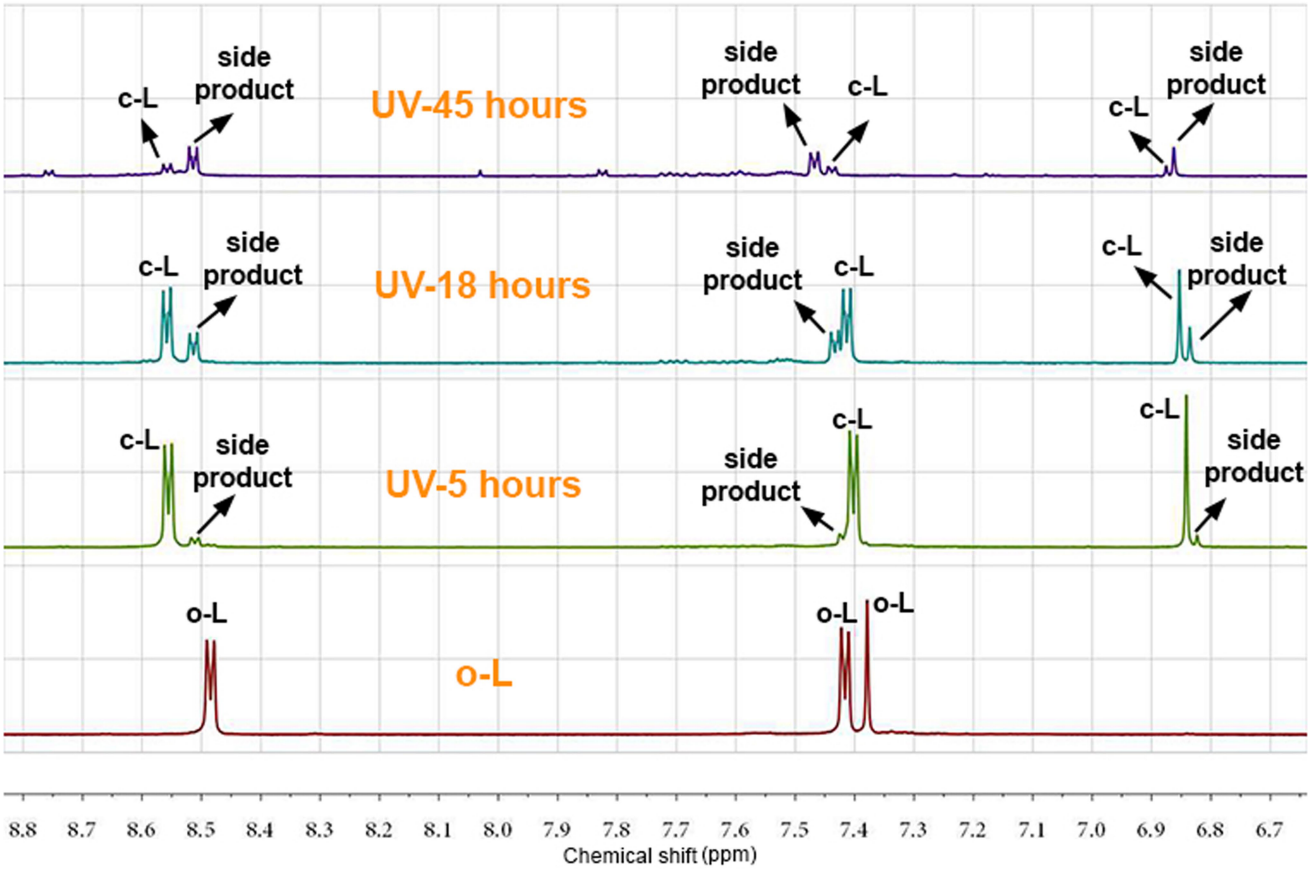
Extended Data Fig. 6 | Simulated results for average connections per MOC for a series of polyMOCs with various Pd_xL_y stoichiometries. Two types of connections between MOCs are defined: total connections and active connections. These connections correspond to two classical models of elasticity: the affine model (total connections, red curve) and the phantom model (active connections, black curve). Both affine (yellow stars) and phantom (blue stars) models were used to calculate the average branch functionality for the o-gel and c-gel based on measured G', and the results were compared with simulated active connections and total connections. For the o-gel, the phantom-model based experimental calculation agrees well with simulated active connections; for the c-gel, the affine-model based experimental calculation agrees well with simulated total connections. The experimental and simulated results suggest that the o-gel is best described as a phantom network while the c-gel is best described as an affine network.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Simulation studies of fatigue behaviours. a, Simulation results for $\overline{f}_{ac}$ of Pd_3L_6 gel obtained by assuming that a certain fraction of ligand is inactive. **b**, Simulation results for $\overline{f}_{tc}$ of $Pd_{24}L_{48}$ gel obtained by assuming that a certain fraction of ligand is inactive.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | ^{1}H NMR of the aromatic regions of model compound o-L during UV irradiation. The o-L was completely converted to c-L after 5 h, with approximately 2% side product as indicated by the presence of a second set of peaks. After 18 h of UV irradiation, about 20% of c-L was converted to the side product. By 45 h, more than 80% of the c-L had undergone side reaction(s).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0366-x

Direct arylation of strong aliphatic C–H bonds

Ian B. Perry[1,3], Thomas F. Brewer[1,3], Patrick J. Sarver[1], Danielle M. Schultz[2], Daniel A. DiRocco[2] & David W. C. MacMillan[1]*

Despite the widespread success of transition-metal-catalysed cross-coupling methodologies, considerable limitations still exist in reactions at sp^3-hybridized carbon atoms, with most approaches relying on prefunctionalized alkylmetal or bromide coupling partners[1,2]. Although the use of native functional groups (for example, carboxylic acids, alkenes and alcohols) has improved the overall efficiency of such transformations by expanding the range of potential feedstocks[3–5], the direct functionalization of carbon–hydrogen (C–H) bonds—the most abundant moiety in organic molecules—represents a more ideal approach to molecular construction. In recent years, an impressive range of reactions that form C(sp^3)–heteroatom bonds from strong C–H bonds has been reported[6,7]. Additionally, valuable technologies have been developed for the formation of carbon–carbon bonds from the corresponding C(sp^3)–H bonds via substrate-directed transition-metal C–H insertion[8], undirected C–H insertion by captodative rhodium carbenoid complexes[9], or hydrogen atom transfer from weak, hydridic C–H bonds by electrophilic open-shell species[10–14]. Despite these advances, a mild and general platform for the coupling of strong, neutral C(sp^3)–H bonds with aryl electrophiles has not been realized. Here we describe a protocol for the direct C(sp^3) arylation of a diverse set of aliphatic, C–H bond-containing organic frameworks through the combination of light-driven, polyoxometalate-facilitated hydrogen atom transfer and nickel catalysis. This dual-catalytic manifold enables the generation of carbon-centred radicals from strong, neutral C–H bonds, which thereafter act as nucleophiles in nickel-mediated cross-coupling with aryl bromides to afford C(sp^3)–C(sp^2) cross-coupled products. This technology enables unprecedented, single-step access to a broad array of complex, medicinally relevant molecules directly from natural products and chemical feedstocks through functionalization at sites that are unreactive under traditional methods.

Metallaphotoredox catalysis has recently emerged as an effective strategy for C(sp^3)–H functionalization[15]. Specifically, the merger of photoredox-mediated hydrogen atom transfer (HAT) and transition-metal catalysis has delivered several methods for the selective functionalization of activated C–H bonds based on low bond dissociation energies and/or polarity effects (α-heteroatom, benzylic and formyl)[10–14]. Inspired by these studies and a strong oxidant-mediated protocol for C(sp^3)–H arylation[16], we proposed that combining a HAT catalyst capable of generating high-energy carbon-centred radicals from strong, inert C–H bonds with the elementary steps of nickel catalysis (aryl oxidative addition, reductive elimination) would enable the coupling of aliphatic carbon frameworks with a range of aryl bromide coupling partners.

We proposed that polyoxometalates (POMs), many of which possess high-energy excited states able to perform the desired C–H abstraction, would be ideal cocatalysts for the proposed transformation[17]. Of particular interest was the decatungstate anion ($[W_{10}O_{32}]^{4-}$), a POM that has been broadly used as an efficient HAT photocatalyst in various oxygenations, dehydrogenations, conjugate additions and, more recently, fluorinations of strong, unactivated, aliphatic C–H bonds[18–23] with bond dissociation energies of up to 100 kcal mol^{-1} (for cyclohexane, ref. [24]). To our knowledge, the decatungstate anion has not previously been merged with transition-metal cross-couplings, and we hoped that such a combination of catalytic processes would enable access to a considerable breadth of carbon-centred radicals and aryl-functionalized products from abundant feedstocks (Fig. 1). Furthermore, the observed selectivity of decatungstate for the abstraction of electron-rich, sterically accessible C–H bonds[25] combined with the steric preference of nickel-catalysed cross-couplings suggested that our proposed dual-catalytic system could provide site-specific arylation of complex organic frameworks.

A detailed description of our proposed mechanism is illustrated in Fig. 2. Photoexcitation of tetrabutylammonium decatungstate (TBADT, **1**) followed by intersystem crossing would produce the triplet excited state (**2**) (with a lifetime, τ, of 55 ns)[26]. Subsequent hydrogen atom

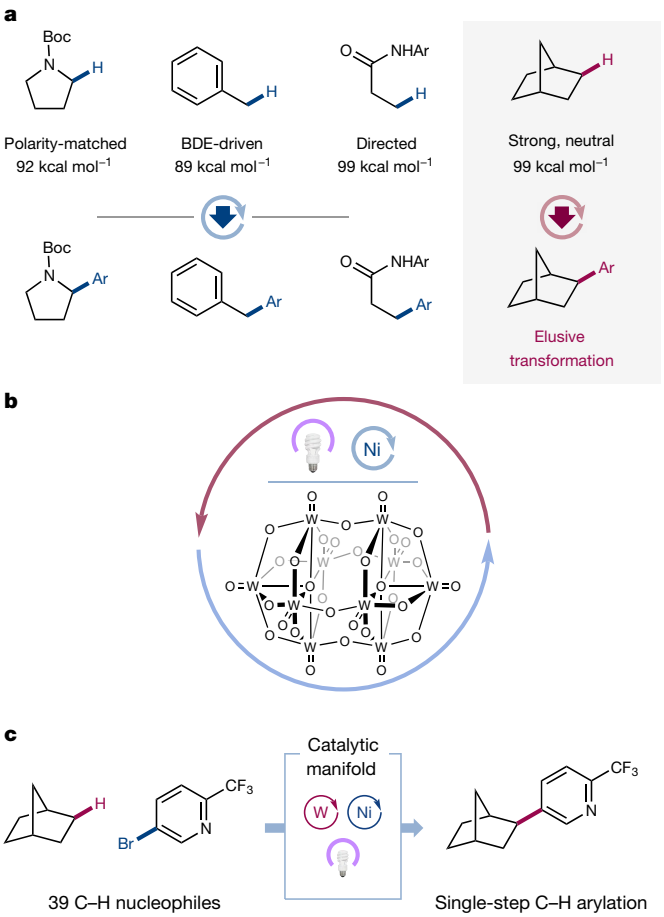


Fig. 1 | Undirected aliphatic C–H arylation. a, Traditional transition-metal-catalysed C(sp^3)–H arylation methods rely on adjacent or distal functionality to facilitate C–H bond activation. **b**, This dual-catalytic approach involves the combination of light-driven, polyoxometalate-facilitated hydrogen atom transfer and nickel catalysis. **c**, Use of this catalytic manifold enables the direct arylation of strong, unactivated C–H bonds. Ar, aryl; BDE, bond dissociation energy; Boc, *tert*-butoxycarbonyl.

[1]Merck Center for Catalysis at Princeton University, Princeton, NJ, USA. [2]Department of Process Chemistry, Merck & Co., Inc., Rahway, NJ, USA. [3]These authors contributed equally: Ian B. Perry, Thomas F. Brewer. *e-mail: dmacmill@princeton.edu

© 2018 Springer Nature Limited. All rights reserved.

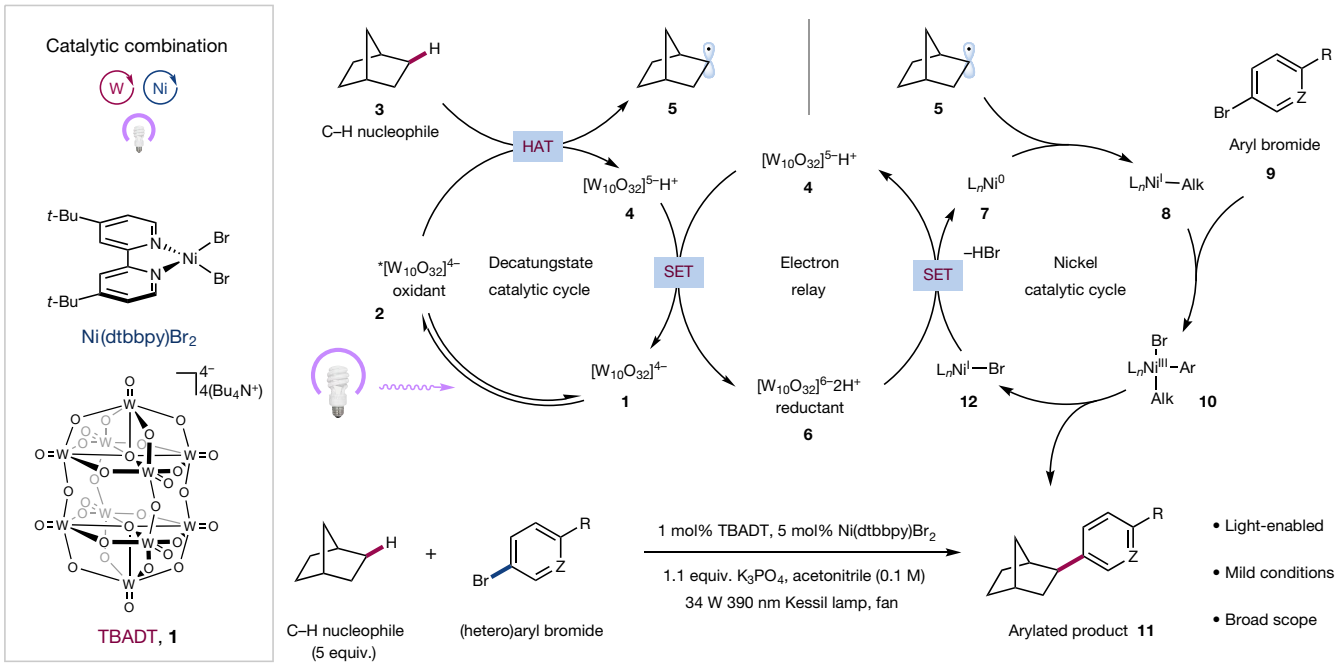


Fig. 2 | Reaction scheme and proposed mechanism for C(sp^3)–H arylation via a dual polyoxometalate HAT and nickel catalytic manifold. The catalytic cycle begins with photoexcitation of the decatungstate anion **1** to provide triplet excited state **2**. HAT from nucleophile **3** affords reduced photocatalyst **4** and open-shell species **5**. Disproportionation of the reduced decatungstate species regenerates the active photocatalyst and affords the reducing hexa-anion **6**. Ni^0 species **7** subsequently captures alkyl radical **5**, furnishing Ni^I-alkyl species **8**. Oxidative addition into aryl electrophile **9** provides Ni^{III} species **10**, which undergoes reductive elimination to afford the product (**11**) and Ni^I–Br species **12**. Single electron transfer (SET) between **6** and **12** regenerates **7** and **4**, closing both catalytic cycles. Alk, alkyl; dtbbpy, 4,4′-di-*tert*-butyl-2,2′-bipyridine; equiv., equivalents; L, ligand; TBADT, tetrabutylammonium decatungstate; *t*-Bu, *tert*-butyl.

abstraction from an alkyl nucleophile such as norbornane (**3**) by excited-state decatungstate (**2**) would readily afford singly reduced decatungstate (**4**) and carbon-centred radical **5**. Disproportionation of singly reduced decatungstate (**4**) would regenerate the active HAT photocatalyst **1** and concurrently form doubly reduced decatungstate (**6**)[26]. Two successive single-electron reductions of precatalyst Ni(dtbbpy)Br_2 (dtbbpy = 4,4′-di-*tert*-butyl-2,2′-bipyridine) (E_p (Ni^{II}/Ni^0) = −1.47 V versus Ag/Ag^+ in acetonitrile, see Supplementary Information) by doubly reduced decatungstate (**6**) ($E_{1/2}^{red}$ ($[W_{10}O_{32}]^{5-}/[W_{10}O_{32}]^{6-}$) = −1.52 V versus Ag/Ag^+ in acetonitrile, see Supplementary Information) could initially afford Ni^0 species **7**, which after capture of alkyl radical **5** would furnish Ni^I-alkyl species **8**. Subsequent oxidative addition into aryl halide **9** by Ni^I-alkyl species **8** would afford Ni^{III}(aryl)(alkyl) species **10**. Reductive elimination would provide the desired cross-coupled product **11** as well as Ni^I species **12**. A final single-electron transfer step between this Ni^I species and the doubly reduced polyoxometalate **6** would regenerate the active Ni^0 catalyst **7**, as well as singly reduced TBADT (**4**), closing both catalytic cycles. An alternative mechanism involving the oxidative addition of Ni^0 catalyst **7** to aryl halide **9** could also be operative[27].

We began our investigation into the proposed transformation by exposing 5-bromo-2-trifluoromethylpyridine and cyclohexane to near-ultraviolet light (Kessil 34 W 390 nm light-emitting diodes (LEDs)) in the presence of the commercially available HAT photocatalyst TBADT, Ni(dtbbpy)Br_2, and potassium phosphate in acetonitrile. To our delight, we observed a 70% analytical yield of the desired cyclohexyl C–H arylation product. Critical to the success of the reaction was the exclusion of both oxygen and water (see Supplementary Information); however, the use of standard benchtop techniques was sufficient in this regard. Moreover, although five equivalents of the C–H nucleophile affords optimal yields, lower substrate loadings can be used albeit with diminished efficiency (see Supplementary Information).

With optimized conditions in hand, we next sought to examine the scope of the transformation with respect to the C–H-bearing partner. As shown in Fig. 3, a diverse array of organic frameworks proved to be competent coupling partners for the C–H arylation protocol. Cycloalkanes with various ring sizes ranging from five to eight carbons were arylated in good yields (**13**–**16**, 57%–70% yield). Linear aliphatic systems were likewise successful in the protocol (**17**–**20**, 41%–56% yield), with a greater-than-statistical preference observed for arylation of the less sterically demanding 2-position for all substrates, including *n*-hexane (**SI-1**, 48% yield, 60% selectivity). Electron-withdrawing substituents further improved this regiocontrol, highlighting the selectivity of decatungstate for more hydridic C–H bonds imparted by the electrophilic nature of its excited state[26]. Accordingly, we found ketones to be particularly effective in modulating regioselectivity, affording products that are functionalized distal to the electron-withdrawing carbonyl moiety (**21**, **22**, **31**–**35**, 31%–65% yield).

This C–H arylation protocol was also found to effectively functionalize a range of electronically diverse primary and secondary benzylic C–H bonds, which were arylated in moderate to good yields (**23**–**25**, 62%–71% yield, see Supplementary Information for three additional examples). Bridged bicyclic alkanes afforded arylated products with complete *exo*-selectivity (**26**–**29** and **35**, 40%–67% yield), probably owing to selective radical capture by nickel catalyst **7** on the less-hindered face. Functionalization of norbornane occurred selectively on the ethylene bridge (**26**, 61% yield). A bromide substituent on the bridging methylene of norbornane was tolerated and, moreover, strongly influenced site-selectivity, giving only the *anti* product (**27**, 67% yield). Notably, heteroatom-containing bicycles afforded the desired products in moderate to good yields (**28** and **29**, 40% and 60% yield, respectively). Arylated lactam **29** was subsequently subjected to ring-opening reductive conditions to afford carbocyclic nucleoside analogue **30** (94% yield), highlighting the utility of the C–H arylation protocol. Intriguingly, adamantane derivatives underwent arylation predominantly at the methylene position (**31** and **32**, 48% and 53% yield, respectively), an unexpected chemoselectivity given that decatungstate-catalysed adamantane functionalization affords 5:1 selectivity for methine positions when corrected for equivalent

© 2018 Springer Nature Limited. All rights reserved.

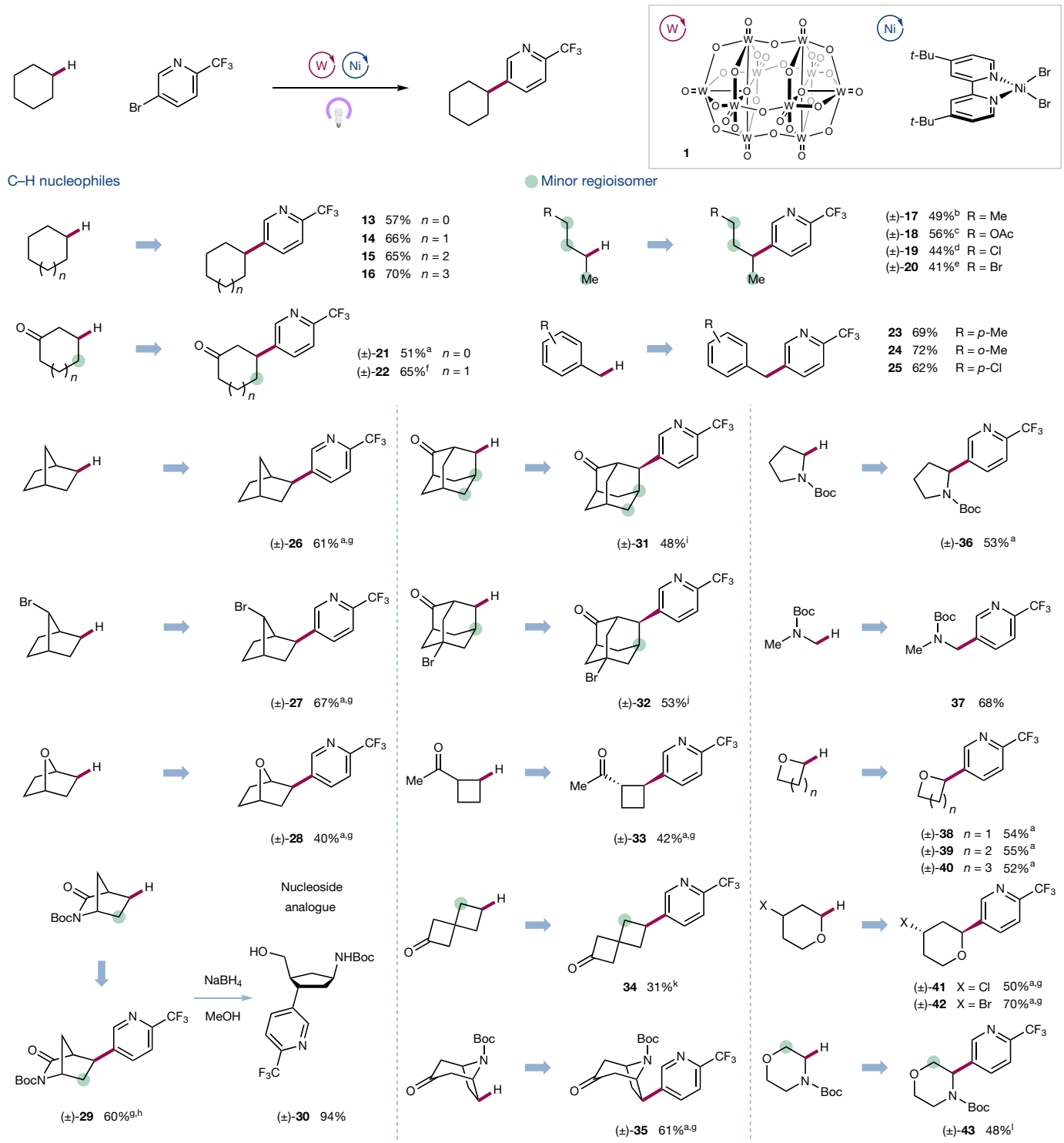


Fig. 3 | Scope of the alkyl nucleophile coupling partner. A broad range of C–H nucleophiles are selectively functionalized by this arylation protocol. Cyclic, acyclic and bicyclic aliphatic systems are amenable substrates. Heteroatom and carbonyl substituents electronically influence regioselectivity, and alkyl halides remain intact. All yields are isolated yields. Conditions as in Fig. 2. Green circles denote sites where notable amounts of other regioisomers are observed. See Supplementary Information for experimental details. Ac, acetyl; Boc, *tert*-butoxycarbonyl; d.r., diastereomeric ratio; Me, methyl; r.r., regioisomeric ratio. [a]>20:1 r.r.; [b]70% selectivity; [c]53% selectivity; [d]79% selectivity; [e]93% selectivity; [f]1.4:1 r.r.; [g]>20:1 d.r.; [h]2.5:1 r.r.; [i]79% selectivity, 3:1 d.r. (major); [j]1.8:1 r.r., 5.4:1 d.r. (major); [k]8.8:1 r.r.; [l]3.4:1 r.r.

hydrogen atoms[28]. This result further highlights the role of the nickel catalyst in determining the regioselectivity of C–C bond formation, presumably via reversible radical capture and selectivity-determining reductive elimination[27]. Four-membered rings were also competent substrates for this arylation protocol, with both an exocyclic ketone and a spirocyclic ketone affording the desired product in moderate yields (**33** and **34**, 42% and 31% yield, respectively). Tropinone, a common scaffold among natural products and pharmaceuticals[29], was also effectively subjected to this dual-catalysis protocol (**35**, 61% yield).

It is important to note that this transformation is not restricted to electronically neutral, unactivated C–H systems. Indeed, various α-heteroatom C–H nucleophiles were readily modified with excellent regioselectivity. As follows from the preference of decatungstate for the most hydridic and sterically accessible C–H bond, *tert*-butoxycarbonyl

© 2018 Springer Nature Limited. All rights reserved.

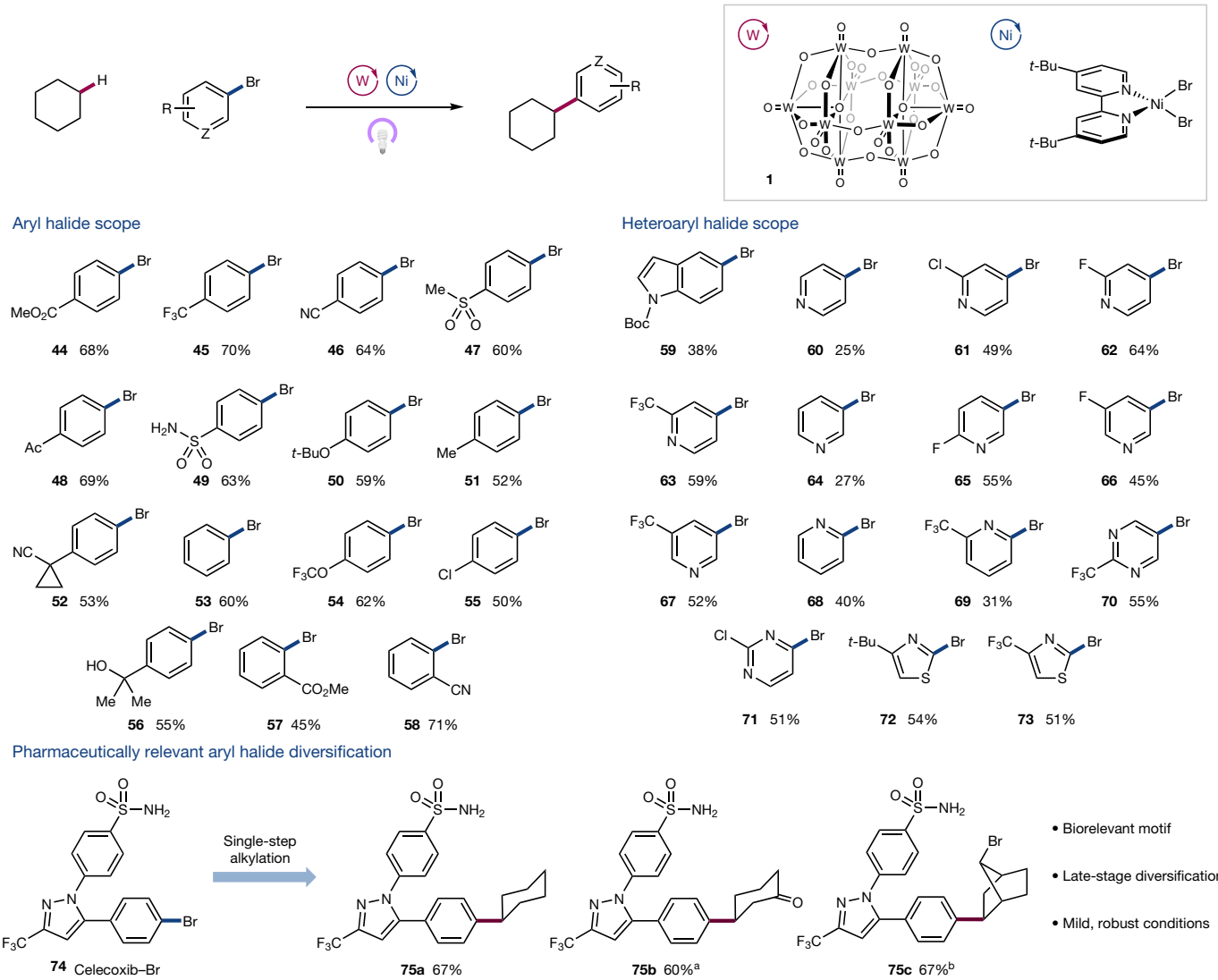


Fig. 4 | Scope of the aryl halide coupling partner. Various electronically diverse aryl bromides were functionalized in moderate to good yields, and unprotected polar functionalities such as alcohols and sulfonamides were well tolerated. Moreover, heteroaryl bromides including indoles, pyridines, pyrimidines and thiazoles were competent coupling partners in the transformation. Finally, the dual catalytic manifold was applied to the synthesis of several analogues of celecoxib. All yields are isolated yields. Conditions as in Fig. 2. See Supplementary Information for experimental details. [a]1.4:1 r.r.; [b]>20:1 d.r.

(Boc)-protected pyrrolidine was functionalized selectively at the α-amino position (**36**, 53% yield). Primary α-amino C–H nucleophile *N*-Boc dimethylamine was also found to be an effective substrate for the transformation (**37**, 68% yield). In addition to nitrogen-containing nucleophiles, various cyclic ethers were regioselectively functionalized in moderate to good yield at the α-oxy position (**38**–**43**, 48%–70% yield). Notable among these substrates, alkyl halides were well tolerated (**41** and **42**, 50 and 70% yield, respectively), opening avenues for subsequent synthetic manipulations. *N*-Boc-morpholine underwent C–H arylation predominantly at the α-amino C–H bond (**43**, 48% yield, 3.4:1 regioisomeric ratio (r.r.)). Useful amounts of the α-oxy product are generated in this case, in contrast to the quinuclidine-mediated triple catalytic arylation reported previously by our laboratory[10] (see Supplementary Information) as well as benzophenone-mediated cyanation[30], wherein exclusive α-amino functionalization is observed.

We next turned our attention to the scope of the aryl halide coupling partner. As shown in Fig. 4, a broad range of electron-deficient aryl bromides provided the desired products in good yield (**44**–**49**, 60%–70% yield). Furthermore, neutral and electron-rich substrates displayed useful coupling efficiencies (**50**–**54**, 52%–62% yield). Chlorine- and fluorine-bearing aryl bromides were alkylated selectively as well (**55** and **SI-5**, 50% and 55% yield, respectively), and free-alcohol-containing substrate **56** was also found to be a competent coupling partner (55% yield). *ortho*-Substituted aryl bromides were likewise alkylated in moderate to good yields (**57** and **58**, 45% and 71% yield, respectively). With respect to heteroaryl bromides, *N*-Boc-indole **59** underwent the desired transformation in useful efficiency (38% yield). A range of bromopyridines were alkylated in useful to good yields as well (**60**–**69**, 25%–64% yield). Bromopyrimidines were effective substrates (**70** and **71**, 55% and 51% yield, respectively), and both electron-rich and electron-deficient 2-bromothiazoles afforded the desired product in moderate yields (**72** and **73**, 54% and 51% yield, respectively). Lastly, the pharmaceutically relevant aryl halide **74**, a precursor to celecoxib was subjected to the reaction conditions with various alkyl C–H nucleophiles. Cyclohexane, cyclohexanone and 7-bromonorbornane were all coupled in good efficiencies (**75a**–**c**, 60%–67% yield), which demonstrates the utility of the protocol in cross-coupling complex aryl fragments with structurally diverse C–H nucleophiles.

Having demonstrated the applicability of the C–H arylation protocol to a broad array of C–H nucleophiles and aryl halide electrophiles, we next investigated its efficacy on naturally occurring aliphatic systems.

© 2018 Springer Nature Limited. All rights reserved.

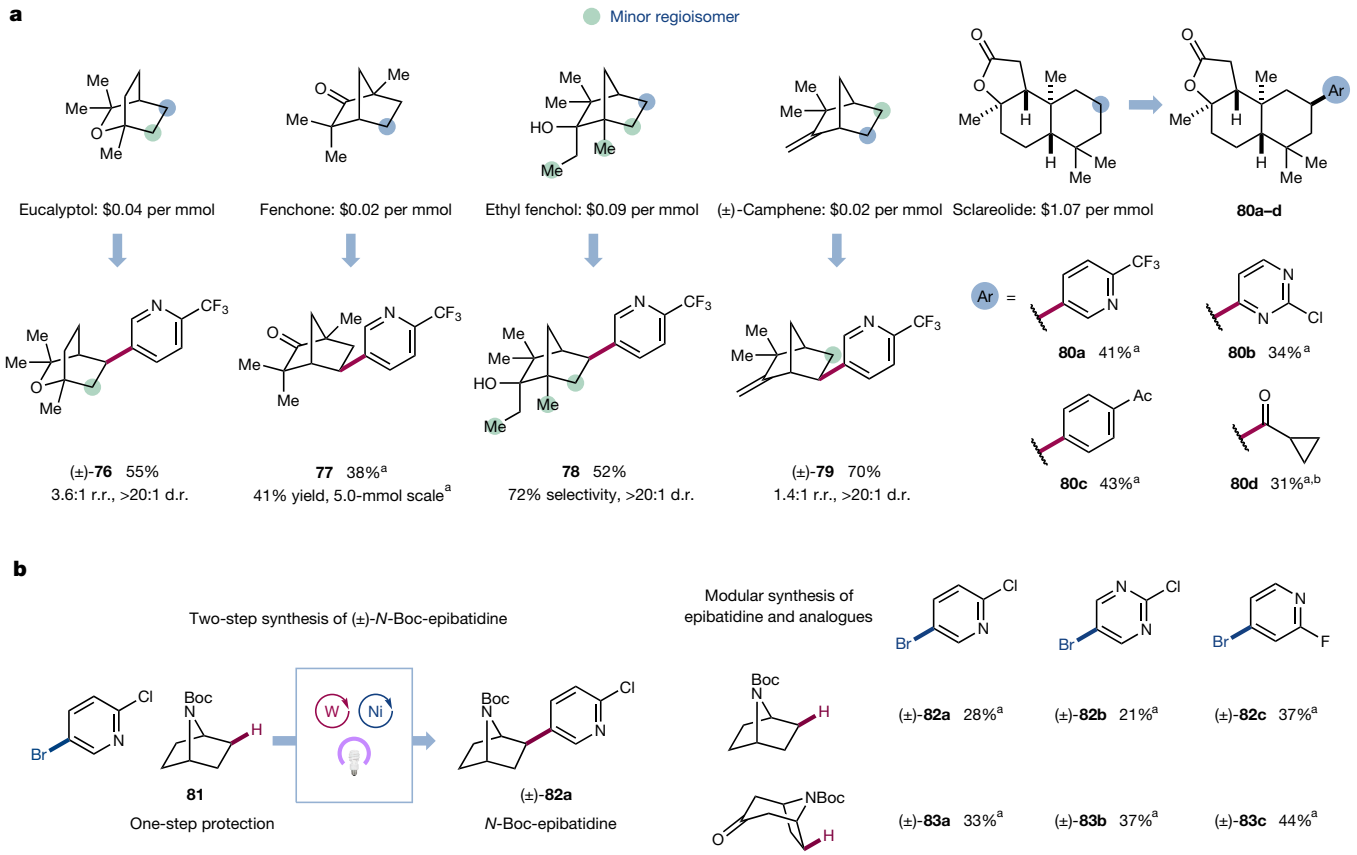


Fig. 5 | Functionalization, synthesis and derivatization of natural products. a, A series of abundantly available terpenoids serve as competent coupling partners in this dual catalytic protocol, affording complex arylated scaffolds in good yields. Major and minor isomers are denoted. **b**, (±)-*N*-Boc-epibatidine was synthesized in two steps from commercially available materials, and subsequently a small library of analogous compounds was constructed. [a]>20:1 r.r., >20:1 d.r.; [b]cyclopropane carbonyl chloride used as electrophilic coupling partner.

As illustrated in Fig. 5a, various inexpensive, abundant natural products were successfully functionalized under our standard conditions, enabling the rapid arylation of complex stereodefined frameworks at carbon sites that lack adaptive functional handles. The observed regioselectivities were in accordance with the expected preferences of this dual-catalytic manifold, as described above. A moderate yield of heteroarene-coupled eucalyptol was observed (**76**, 55% yield), with a strong preference for arylation at the most hydridic and sterically accessible C–H bond. Useful efficiencies were observed for the terpene fenchone, which was also found to be a suitable substrate on 5.0-mmol scale (**77**, 38% and 41% yield, respectively; see Supplementary Information for experimental details). A free-alcohol derivative of fenchone was also readily used in this protocol (**78**, 52% yield), and camphene, a terminal-olefin-containing natural product, was also an effective substrate for arylation (**79**, 70% yield). Lastly, we sought to illustrate the generality of this method by derivatizing the lactone sclareolide with a range of aryl and heteroaryl bromides (**80a–c**, 35%–43% yield). Notably, an alkyl acid chloride provided the sclareolide-derived ketone in useful efficiency (**80d**, 33% yield), which illustrates the capability of our transformation to install a range of functionality onto complex aliphatic substrates without the need for directing groups.

Finally, we demonstrated the application of this C–H arylation protocol towards the rapid generation of complex pharmaceutically relevant molecules. Our target was the natural product epibatidine, a potent non-opioid analgesic. Owing to its high toxicity, epibatidine has seen limited potential as a commercial pharmaceutical[31]; however, a range of epibatidine analogues have been investigated in a clinical setting[32]. We first targeted the synthesis of (±)-*N*-Boc-epibatidine from commercially available 7-azabicyclo[2.2.1]heptane. When *N*-Boc-protected amine substrate **81** and 2-chloro-5-bromopyridine were subjected to the reaction conditions, we observed an unoptimized 28% yield of protected epibatidine (**82a**) in two steps from the commercially available unprotected amine (Fig. 5b, see Supplementary Information for experimental details). To our knowledge, this is the shortest formal synthesis of (±)-epibatidine in the multitude of reported procedures to date[33]. Subsequently, we sought to demonstrate that diversification was possible by variation of both the alkyl fragment and the aryl bromide fragment. A representative sampling of heteroaryl bromides was coupled with bridged bicyclic amines to afford a small set of analogues in synthetically useful yields (**82b** to **83c**, 21%–44% yield). We have developed a robust method for the construction of C(sp^3)–C(sp^2) bonds from alkane nucleophiles and aryl bromide electrophiles. We believe that these results demonstrate the potential to use unactivated C–H bonds as nucleophiles in transition-metal-catalysed cross-coupling transformations.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 20 April 2018; Accepted: 6 June 2018;
Published online 1 August 2018.

1. Jana, R., Pathak, T. P. & Sigman, M. S. Advances in transition metal (Pd, Ni, Fe)-catalyzed cross-coupling reactions using alkyl-organometallics as reaction partners. *Chem. Rev.* **111**, 1417–1492 (2011).
2. Tasker, S. Z., Standley, E. A. & Jamison, T. F. Recent advances in homogeneous nickel catalysis. *Nature* **509**, 299–309 (2014).
3. Zuo, Z. & MacMillan, D. W. C. Decarboxylative arylation of α-amino acids via photoredox catalysis: a one-step conversion of biomass to drug pharmacophore. *J. Am. Chem. Soc.* **136**, 5257–5260 (2014).
4. Lo, J. C., Gui, J., Yabe, Y., Pan, C.-M. & Baran, P. S. Functionalized olefin cross-coupling to construct carbon–carbon bonds. *Nature* **516**, 343–348 (2014).

© 2018 Springer Nature Limited. All rights reserved.

5. Zhang, X. & MacMillan, D. W. C. Alcohols as latent coupling fragments for metallaphotoredox catalysis: sp^3–sp^2 cross-coupling of oxalates with aryl halides. *J. Am. Chem. Soc.* **138**, 13862–13865 (2016).
6. Margrey, K. A., Czaplyski, W. L., Nicewicz, D. A. & Alexanian, E. J. A general strategy for aliphatic C–H functionalization enabled by organic photoredox catalysis. *J. Am. Chem. Soc.* **140**, 4213–4217 (2018).
7. Karimov, R. R. & Hartwig, J. F. Transition-metal-catalyzed selective functionalization of C(sp^3)–H bonds in natural products. *Angew. Chem. Int. Ed.* **57**, 4234–4241 (2018).
8. He, J., Wasa, M., Chan, K. S. L., Shao, Q. & Yu, J.-Q. Palladium-catalyzed transformations of alkyl C–H bonds. *Chem. Rev.* **117**, 8754–8786 (2017).
9. Liao, K., Negretti, S., Musaev, D. G., Bacsa, J. & Davies, H. M. L. Site-selective and stereoselective functionalization of unactivated C–H bonds. *Nature* **533**, 230–234 (2016).
10. Shaw, M. H., Shurtleff, V. W., Terrett, J. A., Cuthbertson, J. D. & MacMillan, D. W. C. Native functionality in triple catalytic cross-coupling: sp^3 C–H bonds as latent nucleophiles. *Science* **352**, 1304–1308 (2016).
11. Le, C., Liang, Y., Evans, R. W., Li, X. & MacMillan, D. W. C. Selective sp^3 C–H alkylation via polarity-match-based cross-coupling. *Nature* **547**, 79–83 (2017).
12. Zhang, X. & MacMillan, D. W. C. Direct aldehyde C–H arylation and alkylation via the combination of nickel, hydrogen atom transfer, and photoredox catalysis. *J. Am. Chem. Soc.* **139**, 11353–11356 (2017).
13. Heitz, D. R., Tellis, J. C. & Molander, G. A. Photochemical nickel-catalyzed C–H arylation: synthetic scope and mechanistic investigations. *J. Am. Chem. Soc.* **138**, 12715–12718 (2016).
14. Shields, B. J. & Doyle, A. G. Direct C(sp^3)–H cross coupling enabled by catalytic generation of chlorine radicals. *J. Am. Chem. Soc.* **138**, 12719–12722 (2016).
15. Twilton, J. et al. The merger of transition metal and photocatalysis. *Nat. Rev. Chem.* **1**, 0052 (2017).
16. Liu, D. et al. Nickel-catalyzed selective oxidative radical cross-coupling: an effective strategy for inert Csp^3–H functionalization. *Org. Lett.* **17**, 998–1001 (2015).
17. Hill, C. L. & Prosser-McCartha, C. M. in *Photosensitization and Photocatalysis Using Inorganic and Organometallic Compounds* (eds Kalyanasundaram, K. & Grätzel, M.) 307–330 (Springer, Dordrecht, 1993).
18. Renneke, R. F. & Hill, C. L. Homogeneous catalytic photochemical functionalization of alkanes by polyoxometalates. *J. Am. Chem. Soc.* **108**, 3528–3529 (1986).
19. Renneke, R. F. & Hill, C. L. Selective photochemical dehydrogenation of saturated hydrocarbons with quantum yields approaching unity. *Angew. Chem. Int. Ed. Engl.* **27**, 1526–1527 (1988).
20. Ravelli, D., Protti, S. & Fagnoni, M. Decatungstate anion for photocatalyzed "window ledge" reactions. *Acc. Chem. Res.* **49**, 2232–2242 (2016).
21. Schultz, D. M. et al. Oxyfunctionalization of the remote C–H bonds of aliphatic amines by decatungstate photocatalysis. *Angew. Chem. Int. Ed.* **56**, 15274–15278 (2017).
22. Halperin, S. D. et al. Development of a direct photocatalytic C–H fluorination for the preparative synthesis of odanacatib. *Org. Lett.* **17**, 5200–5203 (2015).
23. West, J. G., Huang, D. & Sorensen, E. J. Acceptorless dehydrogenation of small molecules through cooperative base metal catalysis. *Nat. Commun.* **6**, 10093 (2015).
24. Luo, Y.-R. *Handbook of Bond Dissociation Energies in Organic Compounds* (CRC Press, Boca Raton, 2003).
25. Ravelli, D., Fagnoni, M., Fukuyama, T., Nishikawa, T. & Ryu, I. Site-selective C–H functionalization by decatungstate anion photocatalysis: synergistic control by polar and steric effects expands the reaction scope. *ACS Catal.* **8**, 701–713 (2018).
26. De Waele, V., Poizat, O., Fagnoni, M., Bagno, A. & Ravelli, D. Unraveling the key features of the reactive state of decatungstate anion in hydrogen atom transfer (HAT) photocatalysis. *ACS Catal.* **6**, 7174–7182 (2016).
27. Gutierrez, O., Tellis, J. C., Primer, D. N., Molander, G. A. & Kozlowski, M. C. Nickel-catalyzed cross-coupling of photoredox-generated radicals: uncovering a general manifold for stereoconvergence in nickel-catalyzed cross-couplings. *J. Am. Chem. Soc.* **137**, 4896–4899 (2015).
28. Ermolenko, L. P. & Giannotti, C. Aerobic photocatalysed oxidation of alkanes in the presence of decatungstates: products and effects of solvent and counter-ion of the catalyst. *J. Chem. Soc. Perkin Trans. 2* **6**, 1205–1210 (1996).
29. Grynkiewicz, G. & Gadzikowska, M. Tropane alkaloids as medicinally useful natural products and their synthetic derivatives as new drugs. *Pharmacol. Rep.* **60**, 439–463 (2008).
30. Hoshikawa, T., Yoshioka, S., Kamijo, S. & Inoue, M. Photoinduced direct cyanation of C(sp^3)–H bonds. *Synthesis* **45**, 874–887 (2013).
31. Bannon, A. W. et al. Broad-spectrum, non-opioid analgesic activity by selective modulation of neuronal nicotinic acetylcholine receptors. *Science* **279**, 77–80 (1998).
32. Badio, B., Garraffo, H. M., Plummer, C. V., Padgett, W. L. & Daly, J. W. Synthesis and nicotinic activity of epiboxidine: an isoxazole analogue of epibatidine. *Eur. J. Pharmacol.* **321**, 189–194 (1997).
33. de Oliveira Filho, R. E. & Omori, A. T. Recent syntheses of frog alkaloid epibatidine. *J. Braz. Chem. Soc.* **26**, 837–850 (2015).

Acknowledgements The research reported here was supported by the National Institutes of Health National Institute of General Medical Sciences (R01 GM103558-03) and gifts from MSD, Bristol-Myers Squibb, Eli Lilly, Genentech, Pfizer and Johnson & Johnson. The authors thank C. Kraml, N. Byrne and L. Wilson (Lotus Separations) for compound purification and I. Pelczer for assistance in structure determination.

Author contributions I.B.P., T.F.B., P.J.S. and D.M.S. performed and analysed the experiments. I.B.P., T.F.B., P.J.S., D.M.S., D.A.D. and D.W.C.M. designed the experiments. I.B.P., T.F.B., P.J.S. and D.W.C.M. prepared the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0366-x.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.W.C.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0309-6

A two-million-year-long hydroclimatic context for hominin evolution in southeastern Africa

Thibaut Caley[1]*, Thomas Extier[1,2], James A. Collins[3,4], Enno Schefuß[5], Lydie Dupont[5], Bruno Malaizé[1], Linda Rossignol[1], Antoine Souron[6], Erin L. McClymont[7], Francisco J. Jimenez-Espejo[8], Carmen García-Comas[9,10], Frédérique Eynaud[1], Philippe Martinez[1], Didier M. Roche[2,11], Stephan J. Jorry[12], Karine Charlier[1], Mélanie Wary[1], Pierre-Yves Gourves[1], Isabelle Billy[1] & Jacques Giraudeau[1]

The past two million years of eastern African climate variability is currently poorly constrained, despite interest in understanding its assumed role in early human evolution[1–4]. Rare palaeoclimate records from northeastern Africa suggest progressively drier conditions[2,5] or a stable hydroclimate[6]. By contrast, records from Lake Malawi in tropical southeastern Africa reveal a trend of a progressively wetter climate over the past 1.3 million years[7,8]. The climatic forcings that controlled these past hydrological changes are also a matter of debate. Some studies suggest a dominant local insolation forcing on hydrological changes[9–11], whereas others infer a potential influence of sea surface temperature changes in the Indian Ocean[8,12,13]. Here we show that the hydroclimate in southeastern Africa (20–25° S) is controlled by interplay between low-latitude insolation forcing (precession and eccentricity) and changes in ice volume at high latitudes. Our results are based on a multiple-proxy reconstruction of hydrological changes in the Limpopo River catchment, combined with a reconstruction of sea surface temperature in the southwestern Indian Ocean for the past 2.14 million years. We find a long-term aridification in the Limpopo catchment between around 1 and 0.6 million years ago, opposite to the hydroclimatic evolution suggested by records from Lake Malawi. Our results, together with evidence of wetting at Lake Malawi, imply that the rainbelt contracted toward the Equator in response to increased ice volume at high latitudes. By reducing the extent of woodland or wetlands in terrestrial ecosystems, the observed changes in the hydroclimate of southeastern Africa—both in terms of its long-term state and marked precessional variability—could have had a role in the evolution of early hominins, particularly in the extinction of *Paranthropus robustus*.

Subtropical southeastern Africa is a region of critical interest because it contains hominin fossils that enable a comparison between continental indicators of hominin evolution and nearby marine records of past climate changes. Different modes of climate change have previously been proposed as major factors that influenced hominin speciation, adaptation or extinction. Some authors stress the effect of long-term trends towards aridity on hominin evolution[1,2] whereas others suggest a crucial role of short periods of extreme climatic variability in driving hominin evolution[3,4].

P. robustus fossils have been found only in southeastern Africa and exclusively in the Limpopo River catchment, at the sites of Cooper's Cave D, Drimolen, Swartkrans, Sterkfontein, Kromdraai B and Gondolin, (Fig. 1) which date from at least about 2 million years ago (Ma) to 0.9 Ma (Extended Data Table 1). It is currently unclear whether climate stress could have had a role in the extinction of this species.

To investigate the hydroclimatic context of the environment in which *P. robustus* lived, we reconstructed changes in the Quaternary hydrological cycle in subtropical southeastern Africa (20–25° S) to determine the drivers of variability and to identify the long-term climate evolution of this region. We used marine sediment core MD96-2048 (26° 10′ 482″ S, 34° 01′ 148″ E, 660-m water depth) from offshore of the Limpopo River mouth (Fig. 1). The chronology of MD96-2048 is established by tuning the $\delta^{18}O$ benthic foraminifera signal to the reference LR04 stack (Fig. 2, Methods), which confirms that the core covers the past 2.14 million years (Myr). We present a multiple-proxy record of hydrological changes in the Limpopo catchment together with a sea surface temperature (SST) record of the southwestern Indian Ocean (Fig. 2).

Modern precipitation in the Limpopo catchment is dominated by austral summer rainfall associated with the extension of the Intertropical Convergence Zone (ITCZ) southwards to 15–20° S (Fig. 1). Changes in the hydrological cycle in the catchment are imprinted on discharge from the Limpopo River. We observe large changes in sedimentary elemental ratios of terrestrial iron to marine calcium, which indicate changes in terrestrial discharge by the Limpopo River at orbital and longer timescales (Fig. 2, Extended Data Fig. 1). Maxima in ln(Fe/Ca) ratios are associated with a more-depleted stable hydrogen isotope composition of plant waxes (δD_{wax}) (Fig. 2), which reflects the isotopic composition of precipitation and is indicative of a higher amount of regional rainfall[10,12,14] (Extended Data Fig. 2, Methods). Maxima in ln(Fe/Ca) ratios are also associated with maxima in concentrations of branched glycerol dialkyl glycerol tetraethers (brGDGT), which are commonly found in soils and attributed to Limpopo River runoff[15], and more enriched plant-wax $\delta^{13}C$. A previous study of the past 0.8 Myr of core MD96-2048 interpreted shifts towards more-depleted $\delta^{13}C_{wax}$ as potentially reflecting more-humid conditions[16]. However, $\delta^{13}C_{wax}$ is a proxy for the contribution of waxes from C_3 versus C_4 plants to the sediments, which can be influenced by many other factors than aridity and humidity. We attribute the enriched plant $\delta^{13}C_{wax}$ at times of depleted δD_{wax} values, and increased ln(Fe/Ca) and concentrations of brGDGT, to stronger transport of C_4 plant material from the upper Limpopo catchment, in addition to the extension of riverine swamps and floodplains that contain abundant C_4 sedges (Fig. 2, Extended Data Figs. 3, 4).

The ln(Fe/Ca) ratio has the highest temporal resolution—300 years on average—of the proxies that we used. Statistical analyses indicate significant 19- and 23-thousand-year (kyr) cycles (precession) and 100-kyr and 400-kyr cycles (eccentricity) but no significant 41-kyr cycle (obliquity) (Extended Data Fig. 5). The dominance of eccentricity and

[1]EPOC, UMR 5805, CNRS, University of Bordeaux, Pessac, France. [2]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France. [3]GFZ – German Research Center for Geosciences, Section 5.1 Geomorphology, Organic Surface Geochemistry Laboratory, Potsdam, Germany. [4]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany. [5]MARUM – Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [6]PACEA, UMR 5199, CNRS, University of Bordeaux, Pessac, France. [7]Department of Geography, Durham University, Durham, UK. [8]Department of Biogeochemistry (JAMSTEC), Yokosuka, Japan. [9]Research and Development Center for Global Change, (JAMSTEC), Yokohama, Japan. [10]Ecology Group, University of Vic – Central University of Catalonia, Barcelona, Spain. [11]Vrije Universiteit Amsterdam, Faculty of Science, Cluster Earth and Climate, Amsterdam, The Netherlands. [12]Unité Géosciences Marines, Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER), Plouzané, France.
*e-mail: thibaut.caley@u-bordeaux.fr

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

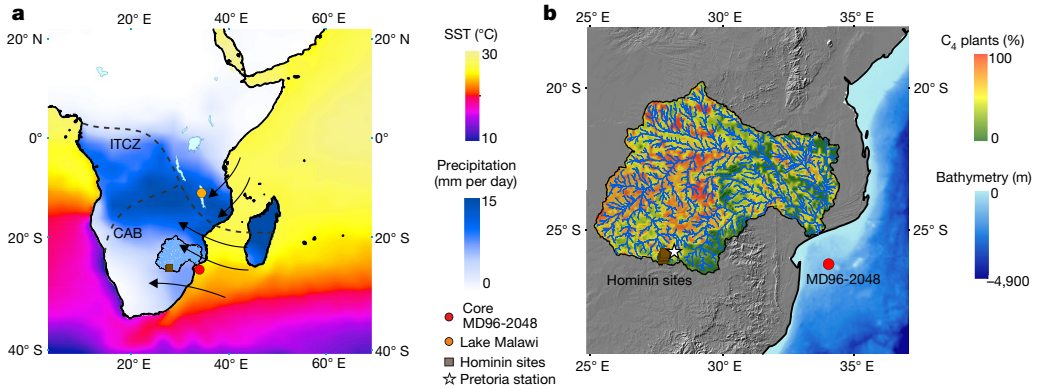


Fig. 1 | Modern climatology over southern Africa and vegetation types in the Limpopo catchment. a, Averaged precipitation rates for January[28] and annual SST over the Indian Ocean[29]. Black arrows represent the atmospheric circulation over southern Africa during austral summer. The ITCZ and the Congo air boundary (CAB) are indicated. **b**, Modelled relative C_4 plant abundance in the Limpopo catchment[30] with indications of topography (from SRTM, Shuttle Radar Topography Mission; https://www2.jpl.nasa.gov/srtm/) and bathymetry (from GEBCO, General Bathymetric Chart of the Oceans; https://www.gebco.net). Location of core MD96-2048, the Lake Malawi records, the main sites of hominin finds for *P. robustus* (sites of Cooper's Cave D, Drimolen, Swartkrans, Sterkfontein, Kromdraai B and Gondolin) and *A. sediba* (Malapa) (Extended Data Table 1) and the Pretoria Global Network of Isotopes in Precipitation station are indicated.

precession cycles indicates a strong influence of low-latitude insolation on hydrological changes in the Limpopo catchment. Rainfall and precession maxima are in phase (that is, maxima of local insolation in the Southern Hemisphere) (Extended Data Figs. 3, 5).

Modern-day rainfall in the subtropical Limpopo region depends mainly on easterly waves and low-pressure cells that largely control summer rainfall (November to March, 72% of the rainfall in Pretoria; Extended Data Fig. 2), and on tropical–extratropical cloud bands and associated thunderstorms[17]. Convective rains are usually associated with the ITCZ and warm, humid easterly winds[17]. Numerical model experiments suggest southern summer insolation forcing exerts a strong and positive effect on monsoon rainfall[11,18]. During precession maxima, higher levels of Southern Hemisphere summer insolation cause higher temperatures and lower surface pressure over the Southern Hemisphere, in particular over land[11]. The contrast between land and ocean temperatures results in stronger easterly moisture inflow into southeastern Africa. Increased rainfall results from increased convection over, and increased humidity transport into, southeastern Africa[11]. Because eccentricity modulates precession amplitudes, the increased summer insolation associated with eccentricity increases the variability in rainfall and fluvial discharge in the Limpopo catchment (Fig. 3).

In addition to this, it is thought that SST anomalies in the Indian Ocean have an influence on summer rainfall in the region[19]. To explore the potential relationship between hydrological cycle changes and oceanic conditions, we reconstructed SSTs for the southwestern Indian Ocean using two different methods (Extended Data Fig. 6, Methods). There is a significant correlation between SST and orbital parameters at the 100-kyr cycle (glacial–interglacial periodicity) and the 41-kyr cycle (obliquity) (Extended Data Fig. 5). The results confirm a previous study that revealed the absence of significant precessional variability in the SST record over the past 0.8 Myr[15]. This suggests that orbital-scale precipitation changes in southeastern Africa are more closely related to the contrast between land and ocean temperatures than to SST changes.

Superimposed on the orbital-scale changes, our record displays a long-term trend towards more arid conditions in southeastern Africa between about 1 and 0.6 Ma (Figs. 2, 3b, c). This period corresponds to the mid-Pleistocene transition (MPT), which is marked by ice-sheet expansion and global SST decrease[20] (Fig. 2). In terms of hydrological changes, the record from Lake Malawi that covers the past 1.3 Myr has previously been interpreted[7,8] to show changes that are opposite to those that we observe for the Limpopo catchment. At Lake Malawi, the climate changed from a predominantly arid environment between 1.3 and about 1 Ma to generally wetter conditions after about 1 Ma[7,8]. This opposing pattern in hydrological changes between the Limpopo catchment and Lake Malawi suggests a gradual contraction of tropical rainfall from the Limpopo catchment towards lower latitudes in response to the ice sheet expansion during the MPT. This rainfall shift could be related to increased Antarctic ice volume during the MPT[21,22].

The long-term trend towards wetter conditions at Lake Malawi has been explained by a progressively less-positive Indian Ocean Dipole since approximately 1 Ma. The Indian Ocean Dipole can enhance or reduce the precessional variability in the low-latitude hydroclimate by modifying the Walker circulation over the Indian Ocean, with a diverse response in eastern Africa[23]. A progressively less-positive Indian Ocean Dipole would have generated wetter conditions in southern Africa and increased the precession signal. Although the precessional signal increased over time in the Limpopo record, the observed progressive increase in aridity is contrary to what would be expected from Indian Ocean Dipole forcing alone (Fig. 3).

On the basis of our new records and comparisons with published records for Lake Malawi, we propose that low-latitude insolation forcing (precession and eccentricity) and changes in ice volume at high latitudes were the main drivers of the southeastern African hydroclimate over the past 2 Myr and that SST forcing had a secondary role. Our results also highlight a large regional variability in the southeastern African hydroclimate.

Changes in the hydrological cycle in southeastern Africa were probably one of multiple factors that influenced the dispersal and evolution of human relatives[2,4]. The more-humid conditions observed between about 2 Ma and 1.75 Ma, associated with a maximum in eccentricity forcing, correspond to several occurrences of *P. robustus* (Fig. 3, Extended Data Table 1). *P. robustus* is a species that was tolerant of environmental variability (eurytopic), but multiple lines of evidence suggest that from its earliest to its last occurrence this species preferred the wooded or humid components—dominated by C_3 plants—of environments that were otherwise dominated by C_4, dry-adapted plants. This preference for habitats dominated by C_3 plants is well-corroborated by the data from palaeoecological studies of other contemporaneous animals that indicate that large quantities of C_4 vegetation were available, but that a more-wooded component and water sources were also present[24] (Fig. 3f, Extended Data Table 2, Methods). This more-humid period between about 2 Ma and 1.75 Ma is also characterized by the presence of *Australopithecus sediba* in the Limpopo catchment (so far known only from the Malapa site)[25]. Multiple lines of evidence suggest *A. sediba* also lived in a wooded or humid habitat and had a diet dominated by C_3 plants, within an otherwise rather-open environment dominated by C_4 plants[25].

Our data raise the possibility that increasing long-term aridity associated with multi-millennial-scale changes after 1 Ma, driven by the

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

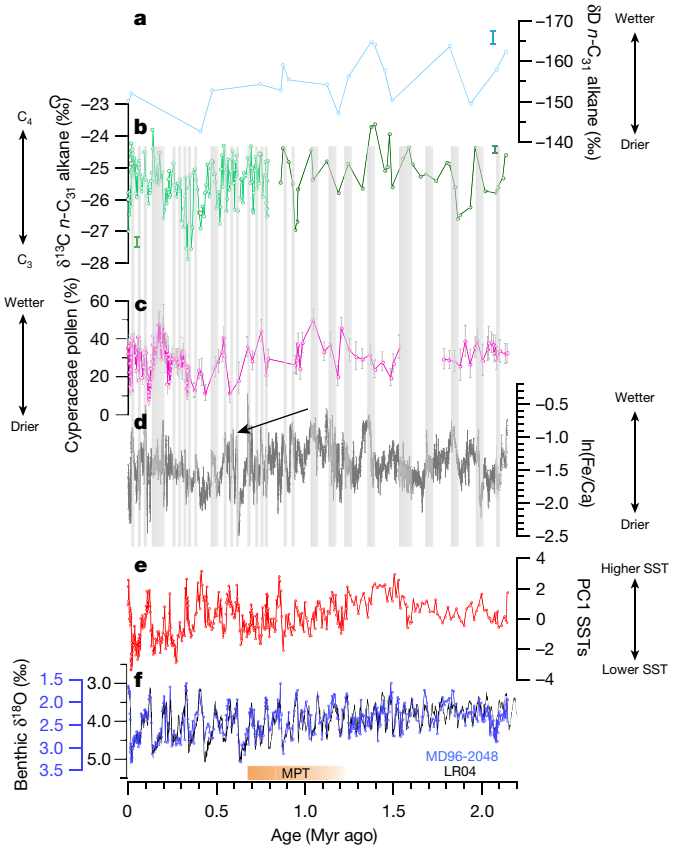


Fig. 2 | Hydrological changes in the Limpopo catchment compared to SSTs of the southwestern Indian Ocean over the past 2.14 Myr. **a**, δD of the *n*-C_{31} alkane of plant waxes. Mean analytical uncertainty of 3‰ is indicated ($n = 8$). **b**, $\delta^{13}C$ of the *n*-C_{31} alkane of plant waxes. Data from a previous study[16] are shown in light green, and data from this study are shown in dark green. Mean analytical uncertainties of 0.2‰ ($n = 15$) and average reproducibility of 0.4‰[16] are indicated. **c**, Pollen percentages of Cyperaceae (data from the past 342 kyr from a previous study[13]). Error bars represent 95% confidence intervals. **d**, ln(Fe/Ca) X-ray fluorescence (XRF) ratios. Arrow indicates the long-term trend discussed in the text. Grey frames represent events of hydrological cycle intensification. **e**, Principal component of the SST records (Methods). **f**, $\delta^{18}O$ of benthic foraminifera compared to the reference LR04 curve (data from the past 790 kyr from a previously study[15]) (Methods). The MPT is indicated.

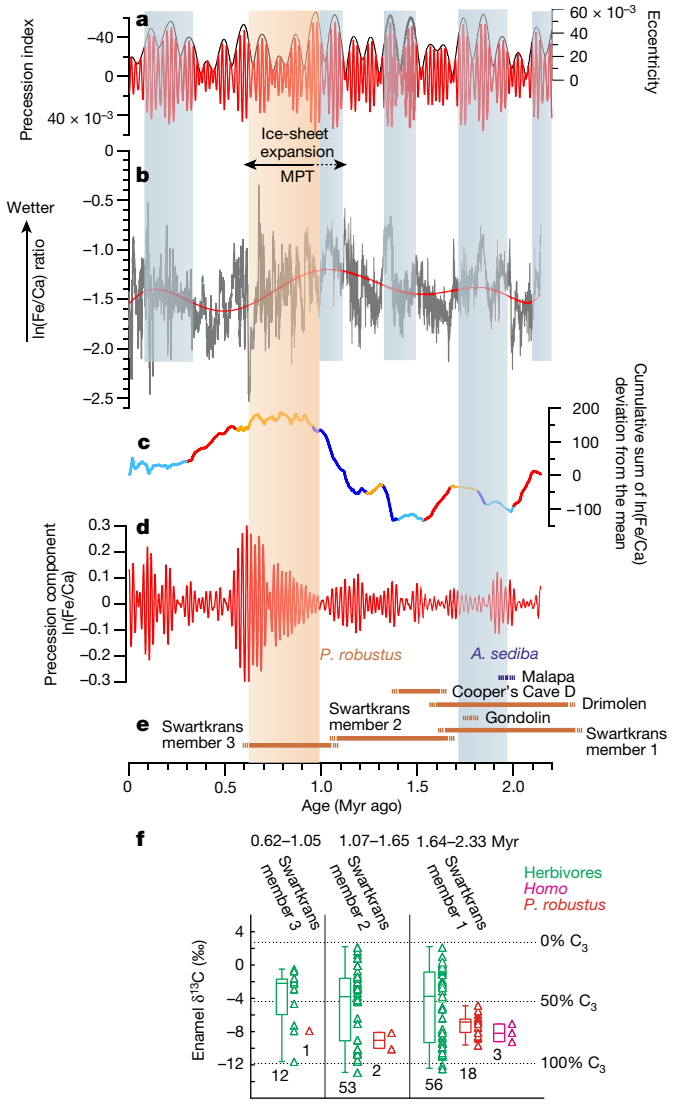


Fig. 3 | Forcings on the hydrological cycle changes in the Limpopo catchment and relationship with hominin evolution over the past 2.14 Myr. **a**, Eccentricity and precession index[31]. **b**, ln(Fe/Ca) XRF ratios as a proxy for hydrological changes in the Limpopo catchment. Red curve denotes a polynomial fit (9th degree). The MPT and associated ice-sheet expansion is indicated (orange shading)[20,21]. Grey shading indicates wetter conditions in the Limpopo catchment, associated with higher eccentricity. **c**, Cumulative sum of ln(Fe/Ca) showing deviation from the mean as indicator of hydrological variability in the Limpopo catchment (Methods) (red curves correspond to dry periods, yellow curves to drying periods, dark blue curves to wet periods and light blue curves to humidification periods). **d**, Precession component of the ln(Fe/Ca) ratios obtained by Gaussian filtering. **e**, Estimated ages for the main sites that yielded remains of the hominins *P. robustus* (Cooper's Cave D, Drimolen, Swartkrans and Gondolin) and *A. sediba* (Malapa) (Extended Data Table 1). **f**, Enamel $\delta^{13}C$ (‰ Vienna Pee Dee Belemnite) of hominins and contemporaneous herbivores at Swartkrans (Extended Data Table 2). Triangles indicate raw data points. The box plots correspond to the median (horizontal line), the interquartile range (box) and the full range of data (vertical whiskers). Sample sizes are indicated by numbers below the plots. Dashed lines highlight the thresholds used to estimate the percentage of C_3 plants-derived foods in the diets.

MPT (Fig. 3b–d), could have diminished the wooded and/or humid component of the habitat preferred by *P. robustus*. This is consistant with a trend towards more open and drier landscapes at Swartkrans (Fig. 3f, Extended Data Table 2, Methods).

It has previously been proposed that extinctions of large mammals are caused mainly by abiotic environmental changes[26]. We propose as a speculative, but plausible, scenario that the geographic ranges of species that preferred wooded and humid habitats—including *P. robustus*—would have contracted and expanded according to precessional (approximately 21-kyr) dry and wet cycles. During the multi-millennial dry periods, population ranges of these species would have contracted and often become fragmented. These isolated populations would have been especially prone to local extinction through lack of sufficient suitable food, water and shelter, and related increased competition and predation. During multi-millennial wet periods associated with precessional maxima, preferred woodland and humid habitats would have expanded again. The surviving populations would have expanded into their previously occupied range, replacing locally extinct populations. The long-term trend toward increased aridity implies that the dry periods became more and more pronounced between 1 and 0.6 Ma (Fig. 3b, d), increasing the likelihood for more numerous extinctions of local populations until the last-remaining population—therefore the species—went extinct[27]. Given that the C_3 component of the vegetation preferred by *P. robustus* was never dominant in the landscape, even during humid periods, populations of this species would have been especially prone to local extinction during dry periods.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Both the long-term state of aridification and the extreme precessional variability in hydroclimate (Fig. 3) could therefore have contributed to the extinction of *P. robustus*.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0309-6.

Received: 20 November 2017; Accepted: 25 April 2018;
Published online 9 July 2018.

1. Vrba, E. S. Environment and evolution: alternative causes of the temporal distribution of evolutionary events. *S. Afr. J. Sci.* **81**, 229–236 (1985).
2. deMenocal, P. B. Plio-Pleistocene African climate. *Science* **270**, 53–59 (1995).
3. Potts, R. Environmental hypotheses of hominin evolution. *Yearb. Phys. Anthropol.* **41**, 93–136 (1998).
4. Maslin, M. A. & Trauth, M. H. in *The First Humans: Origin and Early Evolution of the Genus Homo* (eds Grine, F. E. et al.) 151–158 (Springer, New York, 2009).
5. Cerling, T. E. et al. Woody cover and hominin environments in the past 6 million years. *Nature* **476**, 51–56 (2011).
6. Blumenthal, S. A. et al. Aridity and hominin environments. *Proc. Natl Acad. Sci. USA* **114**, 7331–7336 (2017).
7. Lyons, R. P. et al. Continuous 1.3-million-year record of East African hydroclimate, and implications for patterns of evolution and biodiversity. *Proc. Natl Acad. Sci. USA* **112**, 15568–15573 (2015).
8. Johnson, T. C. et al. A progressively wetter climate in southern East Africa over the past 1.3 million years. *Nature* **537**, 220–224 (2016).
9. Partridge, T. C., Demenocal, P. B., Lorentz, S. A., Paiker, M. J. & Vogel, J. C. Orbital forcing of climate over South Africa: a 200,000-year rainfall record from the Pretoria saltpan. *Quat. Sci. Rev.* **16**, 1125–1133 (1997).
10. Schefuß, E., Kuhlmann, H., Mollenhauer, G., Prange, M. & Pätzold, J. Forcing of wet phases in southeast Africa over the past 17,000 years. *Nature* **480**, 509–512 (2011).
11. Simon, M. H. et al. Eastern South African hydroclimate over the past 270,000 years. *Sci. Rep.* **5**, 18153 (2015).
12. Tierney, J. E. et al. Northern Hemisphere controls on tropical southeast African climate during the past 60,000 years. *Science* **322**, 252–255 (2008).
13. Dupont, L. M. et al. Glacial–interglacial vegetation dynamics in South Eastern Africa coupled to sea surface temperature variations in the Western Indian Ocean. *Clim. Past* **7**, 1209–1224 (2011).
14. Schefuß, E., Schouten, S. & Schneider, R. R. Climatic controls on central African hydrology during the past 20,000 years. *Nature* **437**, 1003–1006 (2005).
15. Caley, T. et al. High-latitude obliquity as a dominant forcing in the Agulhas current system. *Clim. Past* **7**, 1285–1296 (2011).
16. Castañeda, I. S. et al. Middle to Late Pleistocene vegetation and climate change in subtropical southern East Africa. *Earth Planet. Sci. Lett.* **450**, 306–316 (2016).
17. Tyson, P. D. & Preston-Whyte, R. A. *The Weather and Climate of Southern Africa* (Oxford Univ. Press, Oxford, 2000).
18. Kutzbach, J. E., Liu, X., Liu, Z. & Chen, G. Simulation of the evolutionary response of global summer monsoons to orbital forcing over the past 280,000 years. *Clim. Dyn.* **30**, 567–579 (2008).
19. Reason, C. J. C. Subtropical Indian Ocean SST dipole events and southern African rainfall. *Geophys. Res. Lett.* **28**, 2225–2227 (2001).
20. McClymont, E. L., Sosdian, S. M., Rosell-Melé, A. & Rosenthal, Y. Pleistocene sea-surface temperature evolution: early cooling, delayed glacial intensification, and implications for the mid-Pleistocene climate transition. *Earth Sci. Rev.* **123**, 173–193 (2013).
21. Elderfield, H. et al. Evolution of ocean temperature and ice volume through the mid-Pleistocene climate transition. *Science* **337**, 704–709 (2012).
22. Pollard, D. & DeConto, R. M. Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* **458**, 329–332 (2009).
23. Tierney, J. E., Smerdon, J. E., Anchukaitis, K. J. & Seager, R. Multidecadal variability in East African hydroclimate controlled by the Indian Ocean. *Nature* **493**, 389–392 (2013).
24. Balter, V., Braga, J., Télouk, P. & Thackeray, J. F. Evidence for dietary change but not landscape use in South African early hominins. *Nature* **489**, 558–560 (2012).
25. Henry, A. G. et al. The diet of *Australopithecus sediba*. *Nature* **487**, 90–93 (2012).
26. Žliobaitė, I., Fortelius, M. & Stenseth, N. C. Reconciling taxon senescence with the Red Queen's hypothesis. *Nature* **552**, 92–95 (2017).
27. Foley, R. in *African Biogeography, Climate Change, and Human Evolution* (eds Bromage, T. G. & Schrenk, F.) 328–348 (Oxford Univ. Press, Oxford, 1999).
28. Xie, P. & Arkin, P. A. Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates and numerical model outputs. *Bull. Am. Meteorol. Soc.* **78**, 2539–2558 (1997).
29. Locarnini, R. A. et al. *World Ocean Atlas 2009. Volume 1: Temperature* (NOAA Atlas NESDIS 68, U.S. Government Printing Office, Washington, D.C., 2010).
30. Still, C. J. & Powell, R. L. in *Isoscapes: Understanding Movement, Pattern, and Process on Earth through Isotope Mapping* (eds West, J. B. et al.) 179–194 (Springer, Dordrecht, 2010).
31. Laskar, J. et al. A long-term numerical solution for the insolation quantities of the Earth. *Astron. Astrophys.* **428**, 261–285 (2004).

Acknowledgements T.C. is supported by CNRS-INSU. Funding from LEFE-IMAGO CNRS INSU project SeaSalt is acknowledged. T.C. was partly supported by the 'Laboratoire d'Excellence' LabexMER (ANR-10-LABX-19) and co-funded by a grant from the French government under the program 'Investissements d'Avenir', and by a grant from the Regional Council of Brittany (SAD programme). J.A.C. acknowledges funding from the ERC project 'STEEPClim'. E.S. and L.D. acknowledge funding through the DFG Research Center/ Cluster of Excellence 'The Ocean in the Earth System' at MARUM – Center for Environmental Sciences. A.S. acknowledges funding through the LaScArBx, a programme supported by the Agence Nationale de la Recherche (ANR-10-LABX-52). C.G.-C. was supported by CREST (grant number JPMJCR12A3; P.I. SLS) funded by the Japan Science and Technology (JST). Core MD96-2048 was collected during the MOZAPHARE cruise of the RV Marion Dufresne, supported by the French agencies Ministère de l'Education Nationale de la Recherche et de la Technologie, Centre National de la Recherche Scientifique (CNRS) and Institut Paul Emile Victor (IPEV).

Reviewer information *Nature* thanks C. O'Brien, M. Petraglia, K. Uno and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.C. designed the study. T.C., T.E., M.W. and P.-Y.G. performed the Mg/Ca measurements. L.R. analysed the foraminifera assemblages, and T.C. and F.E. analysed the results and performed the transfer function. T.C., T.E., B.M. and K.C. performed the $\delta^{18}O$ analyses on foraminifera. T.C., J.G., P.M. and I.B. performed the XRF measurements and F.J.J.-E. and C.G.-C. conducted the statistical analyses on XRF. J.A.C. and E.S. performed plant-wax δD and $\delta^{13}C$ analyses. L.D. performed the pollen analysis. A.S. and T.C. produced the synthesis on the ecology and environments of South African hominins and conducted the comparisons to the marine record. T.C. and D.M.R. performed and analysed the iLOVECLIM model results. T.C. analysed the results and all authors participated in the interpretation. T.C. wrote the manuscript with contributions from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0309-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

XRF measurements. Element intensities were measured using a XRF-Avaatech core scanner at EPOC. Before analysis, the sediment surface was flattened and covered with Ultralene film. The core sections were scanned at a 0.5-cm resolution at two different levels of energy (10 and 30 keV).

To easily identify periods in XRF ln(Fe/Ca) record, we computed the cumulative sum (cusum) of the deviations from the mean: cusum = sum(ln(Fe/Ca) − (mean(ln(Fe/Ca)). The cumulative sum method was developed for industrial control to detect changes in sequential production[32]. More recently, the cumulative sum has been extensively applied in biological oceanography (for example, ref. [33]) following ref. [34]. We first confirmed the regularity of our time series—no periods with greater sampling effort; the average sampling effort was 283 years with only 8% of scattered time gaps greater than 500 years—and ensured that our periodicity matched that of regularized time series at 500-year and 1,000-year windows. The cumulative sum shows periods (that is, linear sequences) and their value with respect to the long-term average: a positive slope shows a period of values greater than the long-term average, and a negative slope shows a period of values smaller than the long-term average. The steepness of the slope reflects how different a period is from the long-term average. Furthermore, changes in the tendency (that is, sequential periodical changes in the slope) reflect periodical changes from a set of conditions to another (for example, a change from a positive slope to a flat slope in the XRF ln(Fe/Ca) record reflects a change from a humid period to a drier period). In the case of the 1 –0.6-Ma interval, this corresponds to a period of aridification because it is more arid than the previous period between 1.2 Ma and 1 Ma (that is, there is a change of slope from steeply positive to slightly negative). In addition, the period between 1 Ma and 0.6 Ma is very variable, because the slope is noisy rather than straight. The cumulative sum can be found under the function name local.trend, in the R package Pastecs (package of analysis of space-time ecological series).

Plant-wax δD and $\delta^{13}C$. Plant-wax analyses were carried out at MARUM. Samples were oven dried at 40 °C, homogenized and squalane was added as an internal standard before extraction. Lipids were extracted with a DIONEX accelerated colvent extractor using a 9:1 mixture of dichloromethane to methanol at 100 °C and 1,000 p.s.i. for five minutes, repeated three times. The saturated hydrocarbon fraction was obtained by elution of the dried lipid extract over a silica column with hexane and subsequent elution over $AgNO_3$-coated silica to remove unsaturated hydrocarbons.

Compound-specific stable carbon isotope ($\delta^{13}C$) analyses were carried out using a Thermo Fisher Scientific Trace GC Ultra coupled to a Finnigan MAT 252 isotope-ratio-monitoring mass spectrometer via a combustion interface operated at 1,000 °C. Isotope values were calibrated against external CO_2 reference gas and are reported as parts per thousand (‰) against the Vienna Pee Dee Belemnite (V-PDB) standard. Samples were run in at least duplicate. The internal standards yielded a precision of ≤ 0.3‰. Repeated analysis of an external *n*-alkane standard between samples yielded a root-mean-squared accuracy of 0.1‰ and a standard deviation of on average 0.2‰.

When possible, given the high amount of lipids necessary, compound-specific stable hydrogen isotope (δD) compositions were measured using a Thermo Fisher Scientific Trace GC coupled via a pyrolysis reactor operated at 1,420 °C to a Thermo Fisher MAT 253 isotope-ratio mass spectrometer. δD values were calibrated against external H_2 reference gas, the 3H+ factor was monitored daily (values vary between 6.7 and 6.9); δD values are reported in parts per thousand (‰) versus the Vienna Standard of Mean Ocean Water (VSMOW) standard. The internal standards yielded a precision of 2‰ on average. Repeated analysis of an external *n*-alkane standard between samples yielded a root-mean-squared accuracy of <1‰ and a standard deviation of on average 3‰.

The results of a method to adjust the δD_{wax} record for vegetation and ice-volume changes are given in Extended Data Fig. 4.

Pollen preparation. Samples of 1.5 to 8.5 ml were prepared at MARUM. The volume was measured using water displacement. Samples were decalcified with diluted HCl (~12%) and treated with HF (~40%) to remove silicates. Samples were sieved over a screen to remove particles smaller than 10–12 μm. When necessary the sample was decanted to remove remaining silt. Samples were stored in water, mounted in glycerol and microscopically examined (magnification 400× and 1,000×) for pollen and spores. Cyperaceae (sedges) pollen percentages were calculated based on the total number of pollen and spores, ranging from 53 to 365, and 95% confidence intervals were calculated according to a previously published method[35].

$\delta^{18}O$ analyses on foraminifera. Specimens of benthic *Planulina wuellerstorfi* foraminifera were picked from the 250–315-μm size fraction. Analyses were carried out by a coupled system Multiprep–Optima of Micromass at EPOC. The automated preparation system (Multiprep) transforms carbonate samples into CO_2 gas by treatment with orthophosphoric acid at a constant temperature of 75 °C. The CO_2 gas samples were then analysed by isotope mass spectrometry (Optima) in comparison with a calibrated reference gas to determine the isotopic ratio $^{18}O/^{16}O$ of the sample. For all stable oxygen isotope measurements a working standard (Burgbrohl CO_2 gas) was used, which was calibrated against the Vienna Pee Dee Belemnite (V-PDB) standard by using the NBS 19 standard. Analytical standard deviation is about 0.05‰ ($\pm 1\sigma$).

The chronology of the core was established by tuning the $\delta^{18}O$ benthic foraminifera signal to the reference LR04 stack[36] with the AnalySeries software[37] and yielded a correlation coefficient of $R = 0.8$ for the past 2.14 Myr. The core is about 36 m long and the sedimentation rate has a mean value of 2 cm per kyr ($\sigma = 0.91$) and is relatively constant.

SSTs reconstruction. *Globigerinoides ruber sensu stricto* were picked within the 250–315-μm size fraction for trace element analyses. Shells were cleaned at EPOC to eliminate contamination by clays and organic matter based on a previously published procedure[38]. An Agilent inductively coupled plasma optical emission spectrometer (ICP-OES) was used for magnesium and calcium analyses following a previously established procedure[39]. Reproducibility obtained from *G. ruber s. s.* on 80 samples from the complete core was better than 6% ($\pm 1\sigma$, pooled standard deviation). All new analyses for this study ($n = 217$) were performed at EPOC. Measured Mg/Ca ratios were converted into temperature values applying a previously established equation[40], yielding a precision of 1.2 °C.

Total assemblages of planktonic foraminifera were analysed at EPOC using an Olympus SZH10 binocular microscope following previously published taxonomy[41,42]. About 300 specimens were counted in each level after splitting with an Otto microsplitter. Relative abundances of species were used to perform quantification of SST after an ecological transfer function[43] developed at EPOC. The method used here is based on the modern analogue techniques[44] running under the R software, using the ReconstMAT script developed by J. Guiot (BIOINDIC package, https://www.eccorev.fr/spip.php?article389). The modern database used is composed of 367 core tops and derived from core tops covering the southern Indian Ocean[45] in the MARGO project. Calculations of past hydrological parameters rely on a weighted average of SST values from the best five modern analogues, with a maximum weight given to the closest analogue in terms of statistical distance (that is, dissimilarity minimum)[43,46]. This method permits the reconstruction of annual SST with a precision of 0.8 °C (Extended Data Fig. 7).

As each proxy has some uncertainty related to the calibration, non-temperature influences and lateral advection[15], we applied empirical orthogonal function analysis[47] to the two SST records for the past 2.14 Myr (Extended Data Fig. 6). The first principal component (PC1) explains 74% of the total variance for the Mg/Ca and foraminifera transfer function records over the past 2.14 Myr. The correlation between SST proxies and PC1 over the past 2.14 Myr is $R = 0.71$.

Climate modelling. To investigate the control on the past δD composition of precipitation in the Limpopo catchment, we analysed the results of a transient run with the intermediate complexity isotope-enabled climate model iLOVECLIM[48–50] over the past 150 kyr[51]. The atmospheric part of the coupled climate model was run at T21 spatial resolution[51] (~5.65° latitude and longitude) and used accelerated forcing (irradiance, greenhouse gases and ice sheets were updated with an acceleration factor 10). Intermediate complexity models experience some weaknesses caused by the spatial resolution and simplified convective physics but have the advantage of efficient computation. iLOVECLIM was previously successfully applied in the Asian monsoon region[51] and in the West African monsoon region[52] to investigate past monsoonal precipitation changes and their links with changes in the isotopic composition of precipitation. For the current study, analysis of the present-day performance of the iLOVECLIM model for the region, together with results over the past 150 kyr, are shown in Extended Data Fig. 2.

Ecology and environment of the hominins *P. robustus* and *A. sediba* in the Limpopo catchment. Building on previous research[53,54], we produce a short synthesis of multiple lines of evidence to reconstruct the ecology and environmental context of the southern African robust australopith *P. robustus*. We also briefly discuss *A. sediba*. Much literature has been devoted to the ecology of the robust australopiths, and particularly to the ecological differences between *P. robustus* and *P. boisei*, another species that is found in eastern Africa. Given that our geographic scope is centred on South Africa, we here focus mostly on *P. robustus*. It remains a matter of debate whether the robust australopiths—*P. robustus, P. boisei* and *P. aethiopicus* (the probable ancestor of *P. boisei*, also found in eastern Africa)—are closely related to each other in a monophyletic genus or whether robust adaptive traits evolved independently by convergence or parallelism in eastern Africa and southern Africa, which would make the *Paranthropus* genus diphyletic[54,55]. Authors who support the diphyletic hypothesis tend to classify the robust australopiths in the genus *Australopithecus*. Until the issue is settled, here we use the genus name *Paranthropus* as a convenient taxonomic label of a grade that groups these three taxa, which share similar robust adaptive features that differ from the more-gracile genera *Australopithecus* and *Homo*.

Craniomandibular and dental morphology. Species of *Paranthropus* are characterized by a suite of craniomandibular and dental morphological characters: reduced incisors and canines, molarized premolars, enlarged molars, thick

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

enamel, enlarged insertion areas for temporal muscles and a robust mandibular corpus[56]. This suite of robust morphological traits has been frequently interpreted as potential adaptations to diets including hard foods[57,58]. This interpretation is partly supported in *P. robustus* by dental microwear data that indicate occasional consumption of hard objects (see below). Other authors have suggested that these characters reflect the ability of *Paranthropus* to chew tough plant matter for prolonged periods[59,60]. This alternative interpretation is supported by dental microwear and stable carbon isotopic data in *P. boisei*[61,62]. This interpretation could apply to *P. robustus* as well, even though this species probably fed on different kinds of plants to those eaten by *P. boisei* (C_4 herbaceous plants in *P. boisei* versus various C_3 and C_4 plants in *P. robustus*, see below).

Potential modern analogues of herbivorous mammals that evolved from omnivorous ancestors that have thick enamel and bunodont molars include the giant panda *Ailuropoda melanoleuca* and the red panda *Ailurus fulgens*, both of which have independently become highly specialized on a diet of tough bamboo leaves that require prolonged chewing and numerous repeated chewing cycles[63,64]. Both examples illustrate phylogenetic inertia, with evolution acting on preexistent morphologies that limit the range of possible adaptive traits in response to functional selective pressure and result in seemingly suboptimal morphologies for the realized diet. A similar argument has been made for *Paranthropus*[53,65]. Therefore, contrary to herbivorous ungulates that evolved long shearing crests and a more lateral mastication to reduce plant matter into small digestible fragments, the giant panda, the red panda and *Paranthropus* evolved alternative solutions combining thick enamel and increased dental surface with prolonged use of high masticatory forces.

It has also been suggested that an increased dental occlusal surface in mammals could be an adaptive trait for feeding on small bites of small-sized food items[66] by increasing the chance of efficiently masticating the food items with a reduced number of chewing cycles, and limiting the wear induced by a strong increased attrition (tooth-to-tooth contact) in addition to the wear induced by abrasion (tooth-to-food contact). Similar adaptive traits, most notably enlarged surfaces of third molars, evolved independently multiple times in African herbivorous suids[67].

Compared to *Paranthropus*, the craniomandibular and dental morphology of *A. sediba* is more gracile and more similar to that found in older species of *Australopithecus* and younger species of *Homo*. It has recently been argued that the morphology of the cranium was unsuited for feeding on hard objects[68], but that the morphology of the mandible was suited for this activity[69].

Postcranial morphology. Very little is known about the postcranial morphology of *P. robustus* but most studies indicate the retention of traits adaptive for arboreal climbing[70,71]. Hand morphology suggests an ability to perform precision gripping during tool-making and tool-using activities[72]. Numerous bone tools are found in Swartkrans member 3 and Drimolen, sites that have yielded numerous specimens of *P. robustus*. Microscopic and macroscopic wear analyses as well as experimental data suggest that these tools were used for digging into termite mounds[73–75]. The scarcity of *Homo* remains compared to those of *P. robustus* in sites bearing bone tools suggests that the latter species is most likely to be the tool maker and user[73–75]. *A. sediba* is known by two partial skeletons that provide much information on its postcranial morphology. The upper limbs and shoulders of this species retained numerous adaptive traits for climbing and suspensory behaviours[76,77].

Microstructure and biomechanics of enamel. A study of the dental microstructure of *P. robustus* indicates that its enamel was decussated (contrary that of *P. boisei*[78]), which is assumed to reflect a capacity to withstand strong and/or prolonged biomechanic constraints during mastication. Other mammals that feed on hard objects, such as hyaenids[79], and mammals that feed on tough vegetation, such as many ungulates, also display decussated enamel[80].

In addition, an experimental study of the behaviour of enamel under various biomechanic constraints suggests that thick enamel could be an adaptive trait to deal with foods that are either hard or tough, or laden with particulates, potentially including both grit and phytoliths[81]. A recent study observed a low frequency of enamel chipping in *P. robustus*, concluding that it was adapted to eating tough vegetation rather than hard foods[82]. This interpretation is at odds with the dental microwear data that indicate at least some consumption of hard objects (see 'Dental microwear'). Alternatively, we argue that it is equally plausible that this low frequency of chipping is related to the specialized decussated microstructure that reinforces the tooth enamel of *P. robustus* at a microscopic scale, making its teeth less prone to chipping and therefore more durable when consuming hard or tough foods.

Enamel biogeochemistry. Numerous data are available regarding the stable carbon and oxygen isotopic compositions (expressed as $\delta^{13}C$ and $\delta^{18}O$, respectively) and the major elements (Sr, Ba and Ca) of the enamel of *P. robustus*, of other hominins and of other contemporaneous animals. The stable carbon isotope ratios enable the quantification of the proportions of food items deriving, directly or indirectly (through one or several trophic levels), from C_3 plants (mostly woody vegetation, but also sedges in humid environments) and C_4 plants (mostly grass and sedges, but also shrubs of the Amaranthaceae and Chenopodiaceae family). The $\delta^{13}C$ values of the enamel of *P. robustus* indicate a diet that was dominated by foods derived from C_3 plants, with a substantial proportion (about 35–40%) of foods derived from C_4 plants (Fig. 3f and Extended Data Table 2). Much effort has been dedicated to identifying the food items that compose this substantial C_4 component in the diet of *P. robustus*[83–86]. The most probable C_4-plant-derived food items include grass leaves and roots, insects such as grasshoppers and termites, and small vertebrates (birds, lizards, rodents and small ungulates) that consume either C_4 plants or invertebrates that eat C_4 plants. Sedges are also considered to be a likely source of C_4 resources, notably through the consumption of their underground storage organs (USOs)[87,88]. A previous study[84], however, has argued that C_4 sedges were rather scattered in South African riverine settings and that not many of them produced large palatable underground storage organs. This study therefore suggested that C_4 sedges represented only a minor food resource for hominins. However, they only studied the sedges from four riverine sites located in the Kruger National Park and it remains unknown whether their conclusions are valid for the whole Limpopo catchment and larger scales.

More important in our opinion is the fact that the diet of *P. robustus* was dominated by C_3-plant-derived food resources, which could include any parts (leaves, stems, fruits, nuts or underground storage organs) of C_3 plants (mostly trees, shrubs and bushes in wooded environments but also abundant sedges in humid environments). There is little data on the enamel $\delta^{13}C$ values of *A. sediba* ($n=2$) but both specimens display low values (−12.1‰ and −12.2‰) that indicate a diet dominated by C_3-plant-derived food resources[25].

Stable oxygen ratios of enamel are related to multiple factors, including behaviour, ecology, diet, physiology and climate[89]. A previous study[90] classified mammals into evaporation sensitive and evaporation insensitive taxa. The $\delta^{18}O$ values in evaporation sensitive taxa increase with aridity. Evaporation sensitive taxa do not drink much and get most of the water they require from the plants they consume. Conversely, $\delta^{18}O$ values of evaporation insensitive taxa track the $\delta^{18}O$ values of surface water that they drink abundantly and frequently[90,91]. Although hominins were not included in this classification scheme, they are probably water-dependent, as all large primates need to drink a lot of water every day. Hominins generally display relatively low $\delta^{18}O$ values on average compared to the rest of the fauna[62,92], suggesting a high water dependence or consumption of plants containing little evaporated water. Those data support the preference of *P. robustus* for wooded and/or humid environments.

Intra-tooth variations of $\delta^{13}C$ and $\delta^{18}O$ also provide information about intra-annual and inter-annual variation in ecology. Intra-tooth $\delta^{13}C$ and $\delta^{18}O$ profiles were measured using laser ablation on four teeth of *P. robustus* from Swartkrans member 1[83]. The mean values of the $\delta^{13}C$ profiles are similar to those of other *P. robustus* individuals sampled previously, and they also reveal substantial intra-tooth variation with ranges varying from 2‰ to 5‰ over periods representing approximately one or two years (inferred from the number of perikymata). Significant positive correlations between the $\delta^{13}C$ and $\delta^{18}O$ in three out of four specimens of *P. robustus* also indicate that they consumed relatively more C_4-plant-derived foods during the dry season than during the wet season[83]. This could suggest an opportunistic sampling of the environment, with a relative consumption of foods derived from C_3 or C_4 plants that depended in part on seasonal or inter-annual climatic differences. These data are also congruent with the identification of *P. robustus* as an evaporation insensitive taxon and are similar to the pattern observed in some extant evaporation insensitive herbivorous ungulates (for example, strong positive correlations of $\delta^{13}C$ and $\delta^{18}O$ profiles in an extant common hippopotamus *Hippopotamus amphibius*[93]).

Investigations of trace elements preserved in the enamel of *P. robustus* have also revealed interesting patterns[24,94]. *P. robustus* is characterized by relatively high Sr/Ca ratios, lower than in grazing ungulates and slightly higher than in carnivores, browsing ungulates and omnivorous cercopithecid monkeys. *Australopithecus africanus* is characterized by even higher Sr/Ca ratios. Both hominins, but especially *A. africanus*, display low Ba/Ca ratios. Altogether, these data indicate that the proportion of animal matter in their diets is likely to have been low. The only extant animals that combine a high Sr/Ca ratio and a low Ba/Ca ratio are the mole rat (*Cryptomys hottentotus*), and to a lesser extent, the common warthog (*Phacochoerus africanus*). Both species consume large amounts of grass roots. This could indicate that grass root consumption was an important aspect of the diet of *P. robustus*, explaining part of the C_4 component of the diet. However, the Sr/Ca and the Ba/Ca ratios of *A. africanus* and *P. robustus* are not as extreme as those found in the mole rat and the common warthog, suggesting that the consumption of grass roots was probably not as important as in those two species. As previously highlighted[94], one must, however, stress that the use of major element concentrations in enamel as an indication of the diet of extant and extinct animals still require further study. One previous study[24] observed Sr/Ca and Ba/Ca ratios for *P. robustus* that are intermediate between those of *A. africanus* and early *Homo*, and similar to those of browsers: these authors argue the diet of *P. robustus* was dominated by woody plants.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Dental microwear. Investigations of dental microwear have also provided useful information for reconstructing the diet of *P. robustus*[61,95–97]. Dental microwear mostly results from a combination of abrasion (food-to-tooth contact) and attrition (tooth-to-tooth contact), both of which result in various microscopic scars left at the surface of the enamel facets. Among hominins, specimens of *P. robustus* are characterized by unique microwear textures that display strongly variable complexities, ranging from low to very high values, and low anisotropies. This pattern, originally observed on a few specimens, was later confirmed by a large-scale study of dental microwear in *P. robustus*, including numerous specimens from various sites[98]. Similarities in dental microwear textures were noted between *P. robustus* and specimens of several extant primates that are known to consume hard objects such as nuts and seeds (grey-cheeked mangabeys, *Lophocebus albigena* or brown capuchins, *Sapajus apella*), indicating at least some consumption of hard objects by *P. robustus*. Dental microwear data available for *A. sediba*—although they are based on only two specimens—also show high complexities, suggesting a potential consumption of hard objects[25].

Stable carbon isotopes of contemporary mammals. Compiled data of $\delta^{13}C$ values for herbivorous mammals that were contemporary and sympatric with *P. robustus* indicate habitats encompassing a mix of C_3 and C_4 plants, suggesting a mix of woodlands and grasslands in all sites that were sampled (data for Swartkrans members 1, 2, and 3[83,84,99,100]; Cooper's Cave D[101]; Gondolin[102]; Fig. 3f and Extended Data Table 2). The enamel $\delta^{13}C$ values of herbivorous mammals contemporaneous with *P. robustus* from all sites for which sufficient data are available display a bimodal distribution (Extended Data Table 2): the main mode indicates that the habitat was dominated by herbivorous mammals consuming mostly C_4 plants, and the secondary mode indicates that a substantial portion of the remaining herbivorous mammals consumed C_3 plants. Mammals displaying a strong C_4 signal in their enamel probably consumed herbaceous plants, including mostly dry-adapted C_4 grasses in open habitats and possibly C_4 sedges in humid habitats. They are usually classified as grazers. Conversely, mammals displaying a strong C_3 signal in their enamel probably fed mainly on woody plants (trees, shrubs or bushes) in woodlands, and possibly on C_3 sedges in humid habitats. They are usually classified as browsers. Complicating factors of palaeodietary reconstructions include the potential consumption of CAM plants (mostly succulents in arid habitats, and epiphytic plants in closed forests) and C_4 dicot woody vegetation (for example, shrubs of the family Amaranthaceae or Chenopodiaceae). However, the classification of mammals using the dichotomy between C_3 browsers and C_4 grazers inferred from $\delta^{13}C$ values is generally confirmed by ecomorphological and dental wear studies[101]. $\delta^{13}C$ data of *P. robustus* are intermediate between the C_4 pole and the C_3 pole, but closer to the latter, indicating a preference of this hominin for the C_3 wooded and/or humid component of its habitat.

The enamel $\delta^{13}C$ data of the herbivorous mammals found at the Malapa site with *A. sediba* indicate an environment dominated by C_4 plants, presumably grasses, and a few species consuming C_3 plants, presumably woody plants[25]. The pattern is therefore similar to that observed in sites containing *P. robustus*, indicating a preference of *A. sediba* for the C_3 component—presumably woodland—of the otherwise C_4-grass-dominated landscapes.

Ecomorphology and community structures. A previous analysis[103] quantified trophic and locomotor adaptive traits of mammals to compare the ecological diversity of modern and past faunal communities, and found evidence that *P. robustus* inhabited mosaic environments that included both woodlands and grasslands, always close to a water source. This study also detected a pattern of more-open habitats throughout the Swartkrans sequence. This observation supports the information provided by $\delta^{13}C$ values of the herbivorous mammals that—based on our data compilation—generally indicate more C_4 plants in the landscape, presumably grasses adapted to open and dry environments (Fig. 3f). Another previous study[104] conducted a correspondence analysis using the relative abundances of the different groups of mammals classified as woodland-adapted and grassland-adapted, on the basis of stable carbon isotopes and uniformitarian comparisons to extant relatives. This study observed that the relative abundance of *P. robustus* follows the relative abundances of woodland-adapted species and is negatively correlated to the relative abundances of grassland-adapted species. An ecomorphological analysis (L. C. Bishop et al., personal communication) of bovid postcrania from Sterkfontein member 5B 'Oldowan infill' concluded that most bovid species were adapted to open grasslands, and reconstructed the environment as dominated by grasslands but with a nearby more-wooded component. Finally, another previous analysis[106] has suggested that the absence of *P. robustus* in Sterkfontein member 5 west could be linked to a drier local habitat without water-dependent species. Overall, these studies indicate that grasslands were important and probably predominant in the environments occupied by *P. robustus*, but always with a more-wooded and humid component nearby. The ecological data gained from multiple lines of evidence suggest that *P. robustus* preferred this woodland and/or humid component.

Combining the multiple lines of evidence. Based on the evidence discussed above, we interpret *P. robustus* as a species that was tolerant of environmental variability (eurytopic), especially in terms of dietary resources, but with a long-lasting preference for the wooded or humid components (characterized by C_3 plants) of environments that were otherwise dominated by C_4 dry-adapted plants. Such a selective feeding behaviour—with scarce components of the vegetation that are over-represented in the diet of an animal—is frequently observed in extant mammals. For example, extant geladas (*Theropithecus gelada*) from the Guassa Plateau in the highlands of Ethiopia rely extensively on forbs (about 38% of annual diet, and up to 61% of monthly diet) although their preferred forbs represent only 8% of ground cover[107]. A variable C_3-plant-dominated diet, including both hard and tough food items, and displaying strong seasonal or inter-annual variations, is supported by the stable carbon and oxygen isotopes, the dental microwear textures and the overall morphological adaptive traits displayed by *P. robustus*. The robust craniomandibular and dental traits appear as a reasonable compromise to efficiently process an extremely diversified diet including numerous tough parts of plants along with some hard foods, and probably a large number of small food items that had to be eaten in small bites, with some exogenous grit particulates adhering (for example, termites or grass roots).

The preference of *P. robustus* for habitats dominated by C_3 plants, either in woodlands or in humid environments, is well corroborated by the data gained from the study of other animals, indicating large quantities of C_4 vegetation but always with a more wooded component and the nearby presence of a water source. Frequent exploitation of the C_4-plant food resources in the more open component of the landscapes is demonstrated by stable carbon isotopes that indicate a substantial, but not dominant, C_4 component in the diet. Prehensile hands and the possible use of bone tools would have enabled *P. robustus* to access a great variety of foods in both wooded and open habitats. *P. robustus* may be best characterized as an ecotonic species, exploiting intermediate habitats in which the edge effect is maximized, enabling it to forage on a maximal variety of foods in both wooded and open habitats within a limited area while retaining access to secure shelters in woodlands. Such an ecology is displayed by the forest hog (*Hylochoerus meinertzhageni*) that strongly depends on wooded humid forests for shelter, food and water, but which also frequently exploits open grasslands for additional plant resources[108]. Data are scarcer for *A. sediba* but the multiple lines of evidence discussed above suggest an even stronger preference for C_3-plant-dominated wooded habitats, when compared to *P. robustus*. It is worth noting that a similar preference for the C_3-plant component, presumably woodlands, in overall C_4-grass-dominated landscapes is also documented in several other hominin species from eastern Africa[109,110].

Extinction of *P. robustus*. It appears likely that the last documented occurrence of *P. robustus* is dated about 0.9 Ma, or even later (see Extended Data Table 1). Indeed, multiple lines of dating evidence point to a young age of Swartkrans member 3, probably around 0.9 Ma but possibly as young as 0.6 Ma. It is also worth noting the frequency of artificial range truncation of extinct taxa, often called the Signor and Lipps effect[111]: owing to the imperfect nature of the fossil record, the extinction date of a particular taxon is most probably more recent than the last fossil occurrence of this taxon. Bearing in mind that this caution was formulated on the basis of the near-continuous and precisely dated marine palaeontological record, the Signor and Lipps effect is even more pertinent when considering extinction dates based on the spatio-temporally biased continental fossil record of terrestrial faunas (see details in ref. [112]). This dating uncertainty most probably applies to *P. robustus* as the known geographic distribution of this species is extremely small. Most of the sites that contain *P. robustus* are located within a circle of about 3-km radius, and only Gondolin is located a bit further north, around 25 km from Swartkrans and Sterkfontein. A circle of 12.5-km radius encompassing all the known occurrences of *P. robustus* would therefore represent a distribution area of about 500 km^2. Large mammals tend to have much larger geographic distributions than 500 km^2. For comparison, the critically endangered eastern gorilla (*Gorilla beringei*)—the extant species of African great ape that has the smallest geographic distribution—is found over an area of about 70,000 km^2. Other great apes have much larger distributions (*Gorilla gorilla*, over 700,000 km^2; *Pan paniscus*: about 156,000 km^2; *Pan troglodytes*, over 2,600,000 km^2; all data are from the International Union for Conservation of Nature website, http://www.iucnredlist.org/). It is therefore most likely that the real distribution area of *P. robustus* was much larger than the area that is currently sampled by the available fossil record, making the true last occurrence of the species unlikely to be sampled, and the real extinction date of the species probably younger than 0.9 Ma.

The fossil record of South Africa indicates that numerous species became extinct during the Pleistocene. However, the paucity of the record after 1.4 Ma seriously hinders our understanding of the pattern and timing of the extinctions. Middle Pleistocene faunas such as those of Florisbad (about 0.25 Ma) are almost entirely composed of species that are similar to the extant species whereas older faunas (for example, Cornelia at around 1 Ma or Elandsfontein at 1–0.6 Ma) contain a substantial proportion of extinct species (see ref. [113]). Neither a precise date nor a time interval (1–0.5 Ma versus later during the Pleistocene) can be

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

constrained for the extinction of those species based on their currently available fossil records.

Our marine record and previously published terrestrial records are indicative of a clear trend through time towards more-open and drier landscapes in the Limpopo catchment. Given that *P. robustus* preferentially thrived in the non-dominant component of its environment (either woodland or humid grassland) characterized by C_3 plants, and that the diet of this species was dominated by C_3-plant-derived food items, including occasional hard objects, we assume that the regional trend towards a more arid hydroclimate after 1 Ma and the marked precessional variability affected the abundance of this species and its resilience to environmental changes. As the regional hydroclimate became drier, the wooded and humid environments favoured by *P. robustus* probably became progressively scarcer, strongly affecting the fitness and survival of populations of mammals that depended on these habitats for food, water and shelter. We propose a speculative, but plausible, scenario inspired by previous theoretical work[27]. According to this scenario, the geographic ranges of taxa adapted to woodlands and humid habitats, including *P. robustus*, contracted and expanded according to the precessional ~21-kyr dry and wet cycles. During multi-millennial dry periods, the range of populations of these taxa contracted and often became fragmented. The resulting isolated populations were especially prone to local extinction through increased competition and predation induced by the lack of sufficient and suitable food, water and shelters. During multi-millennial wet periods, preferred woodland and humid habitats would have expanded again and the surviving populations would have expanded into their previously occupied range, replacing locally extinct populations. The long-term trend to aridity, inferred from our marine record, implies that dry periods became more and more marked through time, increasing the likelihood for local extinction of numerous local populations, until the extinction of the last remaining local population and therefore the extinction of the species. Thus, both long-term state and the extreme precessional changes in hydroclimate could have affected the evolution of *P. robustus*. A recent synthesis of factors involved in the extinction of large mammals—spanning three continents and the whole Cenozoic period—concluded that abiotic changes, such as climatic changes, were key players in the extinctions of species[26].

What of *Homo*? Several authors (reviewed in ref. [114]) have suggested that strong morphological, behavioural and ecological differences between *P. robustus* and the contemporary *Homo* gave an evolutionary advantage to the latter over the former. Given that *Homo* did not go extinct and remained extremely widespread in Africa and beyond, some authors have related these differences to the extinction of *P. robustus*. Notably different relative abundances of these two taxa do suggest they were occupying separate ecological niches in the sites where they co-occur (Swartkrans 96% *P. robustus*, 4% *Homo*; Drimolen 84% *P. robustus* 16% *Homo*, according to previous work[115]). However, these scenarios are speculative and clearly beyond the scope of our paper, which is focused on *P. robustus*. Regardless of any potential evolutionary advantage of *Homo* over *P. robustus*, whether or not *Homo* went extinct locally in the Limpopo catchment during the aridification period is meaningless to the subsequent evolutionary history of the genus *Homo*. Remains of *Homo* are known in other parts of South Africa at around 1 Ma (for example, Elandsfontein[116] and Cornelia-Uitzoek[117]), as well as in other parts of Africa[118,119], which would have made any local extinction counterbalanced by subsequent dispersals from other regions. The survival of *Homo* could plausibly be explained solely through plain contingency, especially as our synthesis as well as recent literature[114] indicates a eurytopic ecology for both *Homo* and *Paranthropus*.

Code availability. The iLOVECLIM source code is based on the LOVECLIM model version 1.2, for which code is accessible at http://www.elic.ucl.ac.be/modx/elic/index.php?id=289. The developments of the iLOVECLIM source code are hosted at http://forge.ipsl.jussieu.fr/ludus, but are not publicly available owing to copyright restrictions. Access can be granted on demand by request to D.M.R. (didier.roche@lsce.ipsl.fr) to those who conduct research in collaboration with the iLOVECLIM users group.

Data availability. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. Source Data for Fig. 2 is provided with the paper.

32. Page, E. S. Continuous inspection schemes. *Biometrika* **41**, 100–115 (1954).
33. García-Comas, C. et al. Zooplankton long-term changes in the NW Mediterranean Sea: decadal periodicity forced by winter hydrographic conditions related to large-scale atmospheric changes? *J. Mar. Syst.* **87**, 216–226 (2011).
34. Ibanez, F., Fromentin, J. M. & Castel, J. Application of the cumulated function to the processing of chronological data in oceanography. *C. R. Acad. Sci. III* **316**, 745–748 (1993).
35. Maher, L. J. Jr. Nomograms for computing 0.95 confidence limits of pollen data. *Rev. Palaeobot. Palynol.* **13**, 85–93 (1972).
36. Lisiecki, L. E. & Raymo, M. E. A Pliocene–Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography* **20**, PA1003 (2005).
37. Paillard, D., Labeyrie, L. D. & Yiou, P. AnalySeries 1.0: Macintosh program performs time-series analysis. *Eos* **77**, 379 (1996).
38. Barker, S., Greaves, M. & Elderfield, H. A study of cleaning procedures used for foraminiferal Mg/Ca paleothermometry. *Geochem. Geophys. Geosyst.* **4**, 8407 (2003).
39. de Villiers, S., Greaves, M. & Elderfield, H. An intensity ratio calibration method for the accurate determination of Mg/Ca and Sr/Ca of marine carbonates by ICP-AES. *Geochem. Geophys. Geosyst.* **3**, 2001GC000169 (2002).
40. Anand, P., Elderfield, H. & Conte, M. H. Calibration of Mg/Ca thermometry in planktonic foraminifera from a sediment trap time series. *Paleoceanography* **18**, 1050 (2003).
41. Hemleben, C., Spindler, M. & Erson, O. R. *Modern Planktonic Foraminifera* (Springer, New York, 1989).
42. Kennett, J. P. & Srinivasan, M. S. *Neogene Planktonic Foraminifera: A Phylogenetic Atlas* (Hutchinson Ross, Stroudsburg, 1983).
43. Guiot, J. & de Vernal, A. in *Proxies in Late Cenozoic Paleoceanography* (eds Hillaire-Marcel, C. & de Vernal, A.) 523–563 (Elsevier, Amsterdam, 2007).
44. Kucera, M. in *Proxies in Late Cenozoic Paleoceanography* (eds Hillaire-Marcel, C. & de Vernal, A.) 213–262 (Elsevier, Amsterdam, 2007).
45. Barrows, T. T. & Juggins, S. Sea-surface temperatures around the Australian margin and Indian Ocean during the Last Glacial Maximum. *Quat. Sci. Rev.* **24**, 1017–1047 (2005).
46. Kucera, M. et al. Reconstruction of sea-surface temperatures from assemblages of planktonic foraminifera: multi-technique approach based on geographically constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans. *Quat. Sci. Rev.* **24**, 951–998 (2005).
47. Von Storch, H. & Zwiers, F. W. *Statistical Analysis in Climate Research* (Cambridge Univ. Press, Cambridge, 1999).
48. Roche, D. M. $\delta^{18}O$ water isotope in the iLOVECLIM model (version 1.0)–part 1: implementation and verification. *Geosci. Model Dev.* **6**, 1481–1491 (2013).
49. Roche, D. M. & Caley, T. $\delta^{18}O$ water isotope in the *i*LOVECLIM model (version 1.0)–part 2: evaluation of model results against observed $\delta^{18}O$ in water samples. *Geosci. Model Dev.* **6**, 1493–1504 (2013).
50. Caley, T. & Roche, D. M. $\delta^{18}O$ water isotope in the *i*LOVECLIM model (version 1.0)–part 3: a palaeo-perspective based on present-day data–model comparison for oxygen stable isotopes in carbonates. *Geosci. Model Dev.* **6**, 1505–1516 (2013).
51. Caley, T., Roche, D. M. & Renssen, H. Orbital Asian summer monsoon dynamics revealed using an isotope-enabled global climate model. *Nat. Commun.* **5**, 5371 (2014).
52. Collins, J. A. et al. Rapid termination of the African Humid Period triggered by northern high-latitude cooling. *Nat. Commun.* **8**, 1372 (2017).
53. Grine, F. E. & Daegling, D. J. Functional morphology, biomechanics and the retrodiction of early hominin diets. *C. R. Palevol* **16**, 613–631 (2017).
54. Wood, B. & Schroer, K. in *Human Paleontology and Prehistory. Contributions in Honor of Yoel Rak* (eds Marom, A. & Hovers, E.) 95–107 (Springer, Cham, 2017).
55. Patterson, D. B., Faith, J. T., Bobe, R. & Wood, B. Regional diversity patterns in African bovids, hyaenids, and felids during the past 3 million years: the role of taphonomic bias and implications for the evolution of *Paranthropus*. *Quat. Sci. Rev.* **96**, 9–22 (2014).
56. Grine, F. E. (ed). *Evolutionary History of the "Robust" Australopithecines* (Aldine de Gruyter, New York, 1988).
57. Strait, D. S. et al. Viewpoints: diet and dietary adaptations in early hominins: the hard food perspective. *Am. J. Phys. Anthropol.* **151**, 339–355 (2013).
58. Smith, A. L. et al. The feeding biomechanics and dietary ecology of *Paranthropus boisei*. *Anat. Rec. (Hoboken)* **298**, 145–167 (2015).
59. Rabenold, D. & Pearson, O. M. Abrasive, silica phytoliths and the evolution of thick molar enamel in primates, with implications for the diet of *Paranthropus boisei*. *PLoS ONE* **6**, e28379 (2011).
60. Scott, J. E., McAbee, K. R., Eastman, M. M. & Ravosa, M. J. Experimental perspective on fallback foods and dietary adaptations in early hominins. *Biol. Lett.* **10**, 20130789 (2014).
61. Ungar, P. S., Grine, F. E. & Teaford, M. F. Dental microwear and diet of the Plio-Pleistocene hominin *Paranthropus boisei*. *PLoS ONE* **3**, e2044 (2008).
62. Cerling, T. E. et al. Diet of *Paranthropus boisei* in the early Pleistocene of East Africa. *Proc. Natl Acad. Sci. USA* **108**, 9337–9341 (2011).
63. King, R. A. Using *Ailuropoda melanoleuca* as a Model Species for Studying the Ecomorphology of *Paranthropus*. *MSc. thesis, Marshall Univ.* (2014).
64. Weng, Z. Y. et al. Giant panda's tooth enamel: structure, mechanical behavior and toughening mechanisms under indentation. *J. Mech. Behav. Biomed. Mater.* **64**, 125–138 (2016).
65. Ungar, P. S. & Hlusko, L. J. The evolutionary path of least resistance. *Science* **353**, 29–30 (2016).
66. Lucas, P. W. *Dental Functional Morphology: How Teeth Work* (Cambridge Univ. Press, Cambridge, 2004).
67. Souron, A. in *Ecology, Conservation and Management of Wild Pigs and Peccaries* (eds Melletti, M. & Meijaard, E.) 29–38 (Cambridge Univ. Press, Cambridge, 2017).
68. Ledogar, J. A. et al. Mechanical evidence that *Australopithecus sediba* was limited in its ability to eat hard foods. *Nat. Commun.* **7**, 10596 (2016).
69. Daegling, D. J., Carlson, K. J., Tafforeau, P., de Ruiter, D. J. & Berger, L. R. Comparative biomechanics of *Australopithecus sediba* mandibles. *J. Hum. Evol.* **100**, 73–86 (2016).
70. Grine, F. E. & Susman, R. L. Radius of *Paranthropus robustus* from member 1, Swartkrans formation, South Africa. *Am. J. Phys. Anthropol.* **84**, 229–248 (1991).
71. Patel, B. A. The hominoid proximal radius: re-interpreting locomotor behaviors in early hominins. *J. Hum. Evol.* **48**, 415–432 (2005).

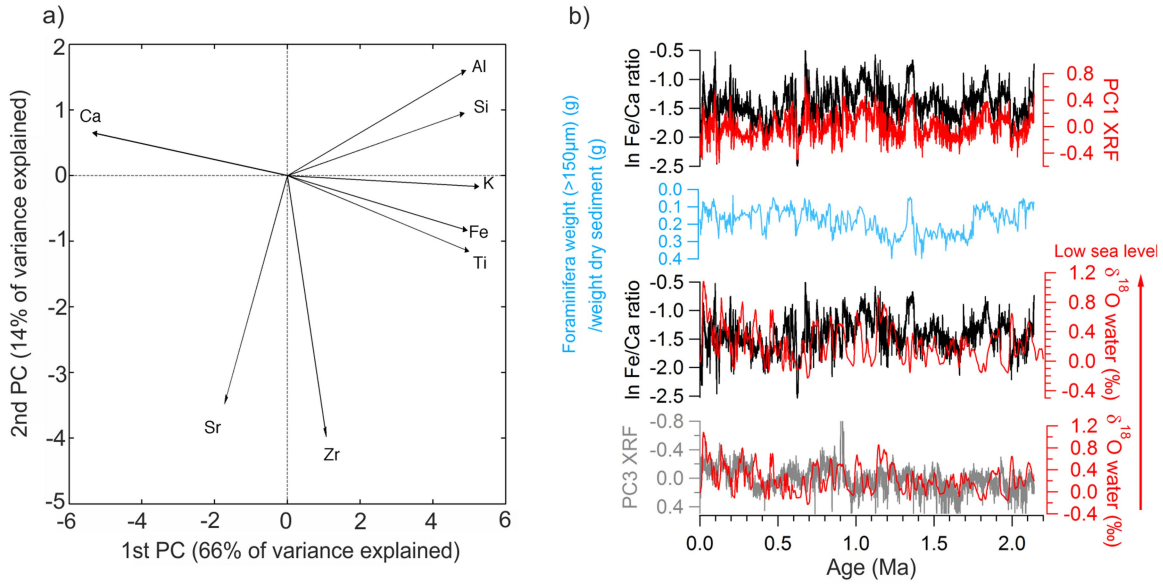
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

72. Susman, R. L. Hand of *Paranthropus robustus* from Member 1, Swartkrans: fossil evidence for tool behavior. *Science* **240**, 781–784 (1988).
73. Backwell, L. R. & d'Errico, F. Additional evidence on the early hominid bone tools from Swartkrans with reference to spatial distribution of lithic and organic artefacts. *S. Afr. J. Sci.* **99**, 259–267 (2003).
74. Backwell, L. R. & d'Errico, F. Early hominid bone tools from Drimolen, South Africa. *J. Archaeol. Sci.* **35**, 2880–2894 (2008).
75. d'Errico, F. & Backwell, L. R. Assessing the function of early hominin bone tools. *J. Archaeol. Sci.* **36**, 1764–1773 (2009).
76. Churchill, S. E. et al. The upper limb of *Australopithecus sediba*. *Science* **340**, 1233477 (2013).
77. Rein, T. R., Harrison, T., Carlson, K. J. & Harvati, K. Adaptation to suspensory locomotion in *Australopithecus sediba*. *J. Hum. Evol.* **104**, 1–12 (2017).
78. Macho, G. in *Trends in Biological Anthropology* Vol. 1 (eds Gerdau-Radonić, K. & McSweeney, K.) 1–10 (Oxbow, Oxford, 2015).
79. Tseng, Z. J. Connecting Hunter–Schreger Band microstructure to enamel microwear features: new insights from durophagous carnivores. *Acta Palaeontol. Pol.* **57**, 473–484 (2012).
80. Alloing-Séguier, L. et al. Enamel microstructure evolution in anthracotheres (Mammalia, Cetartiodactyla) and new insights on hippopotamoid phylogeny. *Zool. J. Linn. Soc.* **171**, 668–695 (2014).
81. Constantino, P. J., Borrero-Lopez, O., Pajares, A. & Lawn, B. R. Simulation of enamel wear for reconstruction of diet and feeding behavior in fossil animals: a micromechanics approach. *BioEssays* **38**, 89–99 (2016).
82. Towle, I., Irish, J. D. & De Groote, I. Behavioral inferences from the high levels of dental chipping in *Homo naledi*. *Am. J. Phys. Anthropol.* **164**, 184–192 (2017).
83. Sponheimer, M. et al. Isotopic evidence for dietary variability in the early hominin *Paranthropus robustus*. *Science* **314**, 980–982 (2006).
84. Sponheimer, M. et al. Hominins, sedges, and termites: new carbon isotope data from the Sterkfontein valley and Kruger National Park. *J. Hum. Evol.* **48**, 301–312 (2005).
85. Sponheimer, M. & Lee-Thorp, J. A. Differential resource utilization by extant great apes and australopithecines: towards solving the C_4 conundrum. *Comp. Biochem. Physiol. A* **136**, 27–34 (2003).
86. Sponheimer, M. in *The Paleobiology of Australopithecus* (eds Reed, K. E., Fleagle, J. G. & Leakey, R. E.) 225–233 (Springer, Dordrecht, 2013).
87. Dominy, N. J., Vogel, E. R., Yeakel, J. D., Constantino, P. & Lucas, P. W. Mechanical properties of plant underground storage organs and implications for dietary models of early hominins. *Evol. Biol.* **35**, 159–175 (2008).
88. Yeakel, J. D., Dominy, N. J., Koch, P. L. & Mangel, M. Functional morphology, stable isotopes, and human evolution: a model of consilience. *Evolution* **68**, 190–203 (2014).
89. Sponheimer, M. & Lee-Thorp, J. A. Oxygen isotopes in enamel carbonate and their ecological significance. *J. Archaeol. Sci.* **26**, 723–728 (1999).
90. Levin, N. E., Cerling, T. E., Passey, B. H., Harris, J. M. & Ehleringer, J. R. A stable isotope aridity index for terrestrial environments. *Proc. Natl Acad. Sci. USA* **103**, 11201–11205 (2006).
91. Faith, J. T. Paleodietary change and its implications for aridity indices derived from $\delta^{18}O$ of herbivore tooth enamel. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **490**, 571–578 (2018).
92. Lee-Thorp, J. A., Sponheimer, M., Passey, B. H., de Ruiter, D. J. & Cerling, T. E. Stable isotopes in fossil hominin tooth enamel suggest a fundamental dietary shift in the Pliocene. *Philos. Trans. R. Soc. Lond. B* **365**, 3389–3396 (2010).
93. Souron, A., Balasse, M. & Boisserie, J.-R. Intra-tooth isotopic profiles of canines from extant *Hippopotamus amphibius* and late Pliocene hippopotamids (Shungura Formation, Ethiopia): insights into the seasonality of diet and climate. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **342–343**, 97–110 (2012).
94. Sponheimer, M., de Ruiter, D., Lee-Thorp, J. & Späth, A. Sr/Ca and early hominin diets revisited: new data from modern and fossil tooth enamel. *J. Hum. Evol.* **48**, 147–156 (2005).
95. Ungar, P. S. & Sponheimer, M. The diets of early hominins. *Science* **334**, 190–193 (2011).
96. Grine, F. E., Sponheimer, M., Ungar, P. S., Lee-Thorp, J. & Teaford, M. F. Dental microwear and stable isotopes inform the paleoecology of extinct hominins. *Am. J. Phys. Anthropol.* **148**, 285–317 (2012).
97. Ungar, P. S., Scott, J. R. & Steininger, C. M. Dental microwear differences between eastern and southern African fossil bovids and hominins. *S. Afr. J. Sci.* **112**, 2015-0393 (2016).
98. Peterson, A. S., Abella, E. F., Grine, F. E., Teaford, M. F. & Ungar, P. S. Microwear textures of *Australopithecus africanus* and *Paranthropus robustus* molars in relation to environment and diet. *J. Hum. Evol.* **119**, 42–63 (2018).
99. Lee-Thorp, J. A., van der Merwe, N. J. & Brain, C. K. Diet of *Australopithecus robustus* at Swartkrans from stable carbon isotopic analysis. *J. Hum. Evol.* **27**, 361–372 (1994).
100. Lee-Thorp, J., Thackeray, J. F. & van der Merwe, N. The hunters and the hunted revisited. *J. Hum. Evol.* **39**, 565–576 (2000).
101. Steininger, C. M. The Dietary Behaviour of Early Pleistocene Bovids from Cooper's Cave and Swartkrans, South Africa. PhD thesis, Univ. of the Witwatersrand (2011).
102. Adams, J. W. Stable carbon isotope analysis of fauna from the Gondolin GD 2 fossil assemblage, South Africa. *Ann. Ditsong Natl. Mus. Nat. Hist.* **2**, 1–5 (2012).
103. Reed, K. E. Early hominid evolution and ecological change through the African Plio-Pleistocene. *J. Hum. Evol.* **32**, 289–322 (1997).
104. de Ruiter, D. J., Sponheimer, M. & Lee-Thorp, J. A. Indications of habitat association of *Australopithecus robustus* in the Bloubank Valley, South Africa. *J. Hum. Evol.* **55**, 1015–1030 (2008).
106. Kuman, K. & Clarke, R. J. Stratigraphy, artefact industries and hominid associations for Sterkfontein, member 5. *J. Hum. Evol.* **38**, 827–847 (2000).
107. Fashing, P. J., Nguyen, N., Venkataraman, V. V. & Kerby, J. T. Gelada feeding ecology in an intact ecosystem at Guassa, Ethiopia: variability over time and implications for theropith and hominin dietary evolution. *Am. J. Phys. Anthropol.* **155**, 1–16 (2014).
108. d'Huart, J.-P. in *Pigs, Peccaries, and Hippos: Status Survey and Conservation Action Plan* (ed. Oliver, W. L. R.) 84–92 (IUCN, Gland, 1993).
109. Quinn, R. L. et al. Pedogenic carbonate stable isotopic evidence for wooded habitat preference of early Pleistocene tool makers in the Turkana Basin. *J. Hum. Evol.* **65**, 65–78 (2013).
110. Robinson, J. R., Rowan, J., Campisano, C. J., Wynn, J. G. & Reed, K. E. Late Pliocene environmental change during the transition from *Australopithecus* to *Homo*. *Nat. Ecol. Evol.* **1**, 0159 (2017).
111. Signor, P. W. III & Lipps, J. H. in *Geological Implications of Impacts of Large Asteroids and Comets on the Earth* (eds Silver, L. T. & Schultz, P. H.) 291–296 (The Geological Society of America, Boulder, 1982).
112. White, T. D. in *Paleoclimate and Evolution, with Emphasis on Human Origins* (eds Vrba, E. S. et al.) 369–384 (Yale Univ. Press, New Haven, 1995).
113. Codron, D., Brink, J. S., Rossouw, L. & Clauss, M. The evolution of ecological specialization in southern African ungulates: competition- or physical environmental turnover? *Oikos* **117**, 344–353 (2008).
114. Wood, B. & Strait, D. Patterns of resource use in early *Homo* and *Paranthropus*. *J. Hum. Evol.* **46**, 119–162 (2004).
115. Moggi-Cecchi, J., Menter, C., Boccone, S. & Keyser, A. Early hominin dental remains from the Plio-Pleistocene site of Drimolen, South Africa. *J. Hum. Evol.* **58**, 374–405 (2010).
116. Klein, R. G., Avery, G., Cruz-Uribe, K. & Steele, T. E. The mammalian fauna associated with an archaic hominin skullcap and later Acheulean artifacts at Elandsfontein, Western Cape Province, South Africa. *J. Hum. Evol.* **52**, 164–186 (2007).
117. Brink, J. S. et al. First hominine remains from a ~1.0 million year old bone bed at Cornelia-Uitzoek, Free State Province, South Africa. *J. Hum. Evol.* **63**, 527–535 (2012).
118. Asfaw, B. et al. Remains of *Homo erectus* from Bouri, Middle Awash, Ethiopia. *Nature* **416**, 317–320 (2002).
119. Abbate, E. et al. A one-million-year-old *Homo* cranium from the Danakil (Afar) Depression of Eritrea. *Nature* **393**, 458–460 (1998).
120. Bahr, A. et al. Deciphering bottom current velocity and paleoclimate signals from contourite deposits in the Gulf of Cádiz during the last 140 kyr: an inorganic geochemical approach. *Geochem. Geophys. Geosyst.* **15**, 3145–3160 (2014).
121. Adegbie, A. T., Schneider, R. R., Röhl, U. & Wefer, G. Glacial millennial-scale fluctuations in central African precipitation recorded in terrigenous sediment supply and freshwater signals offshore Cameroon. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **197**, 323–333 (2003).
122. Dickson, A. J., Leng, M. J., Maslin, M. A. & Röhl, U. Oceanic, atmospheric and ice-sheet forcing of South East Atlantic Ocean productivity and South African monsoon intensity during MIS-12 to 10. *Quat. Sci. Rev.* **29**, 3936–3947 (2010).
123. Revel, M. et al. 20,000 years of Nile River dynamics and environmental changes in the Nile catchment area as inferred from Nile upper continental slope sediments. *Quat. Sci. Rev.* **130**, 200–221 (2015).
124. Ziegler, M. et al. Development of Middle Stone Age innovation linked to rapid climate change. *Nat. Commun.* **4**, 1905 (2013).
125. Rohling, E. J. et al. Sea-level and deep-sea-temperature variability over the past 5.3 million years. *Nature* **508**, 477–482 (2014).
126. IAEA. *Isotope Hydrology Information System, The ISOHIS Database* http://www.iaea.org/water (2006).
127. Risi, C., Bony, S. & Vimeux, F. Influence of convective processes on the isotopic composition ($\delta^{18}O$ and δD) of precipitation and water vapor in the tropics: 2. physical interpretation of the amount effect. *J. Geophys. Res.* **113**, D19306 (2008).
128. Stock, W. D., Chuba, D. K. & Verboom, G. A. Distribution of South African C_3 and C_4 species of Cyperaceae in relation to climate and phylogeny. *Austral Ecol.* **29**, 313–319 (2004).
129. Dupont, L. M. & Kuhlmann, H. Glacial-interglacial vegetation change in the Zambezi catchment. *Quat. Sci. Rev.* **155**, 127–135 (2017).
130. Schrag, D. P. et al. The oxygen isotopic composition of seawater during the Last Glacial Maximum. *Quat. Sci. Rev.* **21**, 331–342 (2002).
131. Sachse, D. et al. Molecular paleohydrology: interpreting the hydrogen-isotopic composition of lipid biomarkers from photosynthesizing organisms. *Annu. Rev. Earth Planet. Sci.* **40**, 221–249 (2012).
132. Collins, J. A. et al. Estimating the hydrogen isotopic composition of past precipitation using leaf-waxes from western Africa. *Quat. Sci. Rev.* **65**, 88–101 (2013).
133. Schefuß, E., Schouten, S., Jansen, J. H. & Sinninghe Damsté, J. S. African vegetation controlled by tropical sea surface temperatures in the mid-Pleistocene period. *Nature* **422**, 418–421 (2003).
134. Grinsted, A., Moore, J. C. & Jevrejeva, S. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Process. Geophys.* **11**, 561–566 (2004).
135. Schulz, M. & Mudelsee, M. REDFIT: estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. *Comput. Geosci.* **28**, 421–426 (2002).
136. Imbrie, J. et al. in *Milankovitch and Climate: Understanding the Response to Astronomical Forcing* (eds Berger, A. et al.) 269–305 (Springer, Dordrecht, 1984).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

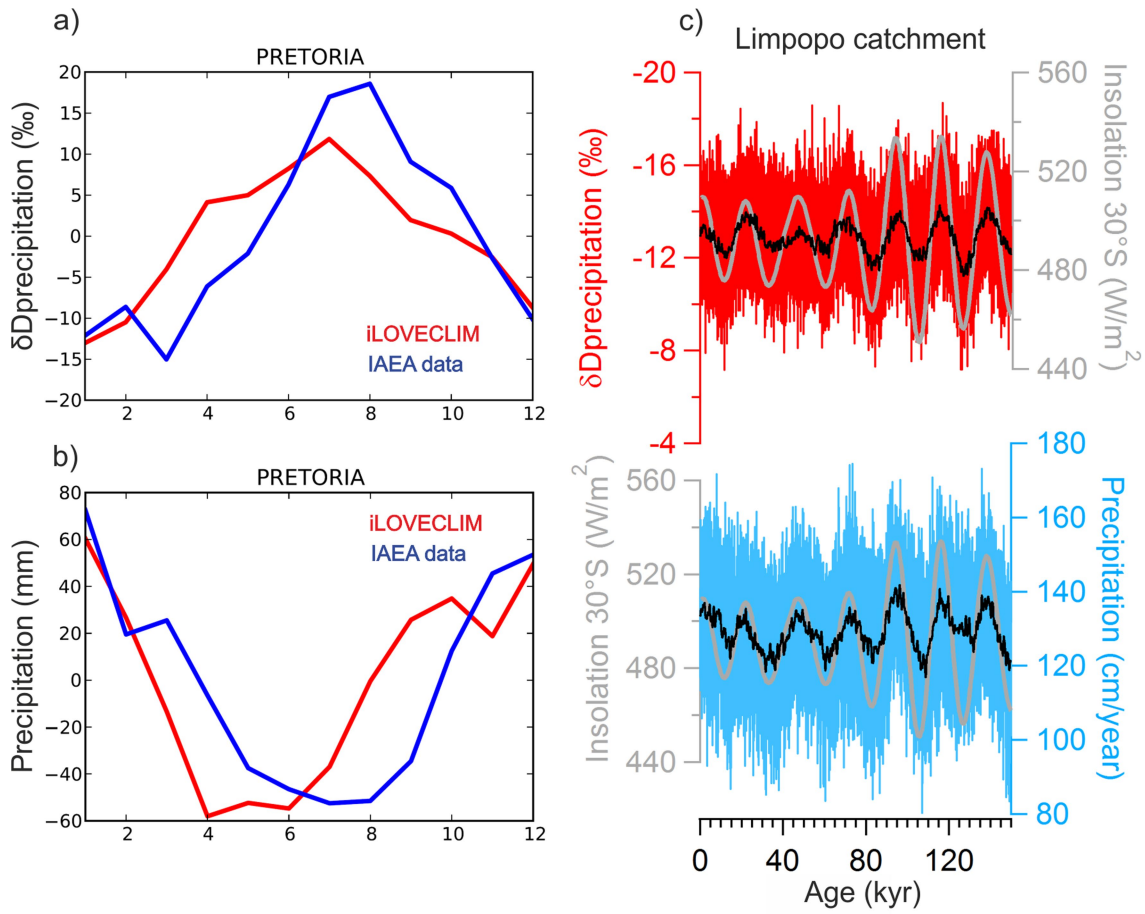
137. Dirks, P. H. et al. Geological setting and age of *Australopithecus sediba* from southern Africa. *Science* **328**, 205–208 (2010).
138. Pickering, R. et al. *Australopithecus sediba* at 1.977 Ma and implications for the origins of the genus *Homo*. *Science* **333**, 1421–1423 (2011).
139. Berger, L. R., de Ruiter, D. J., Steininger, C. M. & Hancox, J. Preliminary results of excavations at the newly investigated Coopers D deposit, Gauteng, South Africa. *S. Afr. J. Sci.* **99**, 276–278 (2003).
140. de Ruiter, D. J. et al. New *Australopithecus robustus* fossils and associated U-Pb dates from Cooper's Cave (Gauteng, South Africa). *J. Hum. Evol.* **56**, 497–513 (2009).
141. Keyser, A. W., Menter, C. G., Moggi-Cecchi, J., Rayne Pickering, T. & Berger, L. R. Drimolen: a new hominid-bearing site in Gauteng, South Africa. *S. Afr. J. Sci.* **96**, 193–197 (2000).
142. Adams, J. W., Rovinsky, D. S., Herries, A. I. R. & Menter, C. G. Macromammalian faunas, biochronology and palaeoecology of the early Pleistocene Main Quarry hominin-bearing deposits of the Drimolen Palaeocave System, South Africa. *PeerJ* **4**, e1941 (2016).
143. Thackeray, J. F., Kirschvink, J. L. & Raub, T. D. Palaeomagnetic analyses of calcified deposits from the Plio-Pleistocene hominid site of Kromdraai, South Africa: news & views. *S. Afr. J. Sci.* **98**, 537–540 (2002).
144. Herries, A. I. R., Curnoe, D. & Adams, J. W. A multi-disciplinary seriation of early *Homo* and *Paranthropus* bearing palaeocaves in southern Africa. *Quat. Int.* **202**, 14–28 (2009).
145. Herries, A. I. R & Adams, J. W. Clarifying the context, dating and age range of the Gondolin hominins and *Paranthropus* in South Africa. *J. Hum. Evol.* **65**, 676–681 (2013).
146. Braga, J., Fourvel, J.-B., Lans, B., Bruxelles, L. & Thackeray, J. F. in *Kromdraai. A Birthplace of* Paranthropus *in the Cradle of Humankind* (eds Braga, J. & Thackeray, J. F.) 1–16 (SUN, Stellenbosch, 2016).
147. Herries, A. I., Adams, J. W., Kuykendall, K. L. & Shaw, J. Speleology and magnetobiostratigraphic chronology of the GD 2 locality of the Gondolin hominin-bearing paleocave deposits, North West Province, South Africa. *J. Hum. Evol.* **51**, 617–631 (2006).
148. Adams, J. W., Herries, A. I., Kuykendall, K. L. & Conroy, G. C. Taphonomy of a South African cave: geological and hydrological influences on the GD 1 fossil assemblage at Gondolin, a Plio-Pleistocene paleocave system in the Northwest Province, South Africa. *Quat. Sci. Rev.* **26**, 2526–2543 (2007).
149. Curnoe, D. K. A. Contribution to the Question of Early *Homo* in Southern Africa: Researches into Dating, Taxonomy and Phylogeny Reconstruction. PhD thesis, Australian National University (1999).
150. Herries, A. I. & Shaw, J. Palaeomagnetic analysis of the Sterkfontein palaeocave deposits: implications for the age of the hominin fossils and stone tool industries. *J. Hum. Evol.* **60**, 523–539 (2011).
151. Granger, D. E. et al. New cosmogenic burial ages for Sterkfontein Member 2 *Australopithecus* and Member 5 Oldowan. *Nature* **522**, 85–88 (2015).
152. Gibbon, R. J. et al. Cosmogenic nuclide burial dating of hominin-bearing Pleistocene cave deposits at Swartkrans, South Africa. *Quat. Geochronol.* **24**, 10–15 (2014).
153. Curnoe, D., Grün, R., Taylor, L. & Thackeray, F. Direct ESR dating of a Pliocene hominin from Swartkrans. *J. Hum. Evol.* **40**, 379–391 (2001).
154. Pickering, R., Kramers, J. D., Hancox, P. J., de Ruiter, D. J. & Woodhead, J. D. Contemporary flowstone development links early hominin bearing cave deposits in South Africa. *Earth Planet. Sci. Lett.* **306**, 23–32 (2011).
155. Balter, V. et al. U–Pb dating of fossil enamel from the Swartkrans Pleistocene hominid site, South Africa. *Earth Planet. Sci. Lett.* **267**, 236–246 (2008).
156. Vrba, E. S. Some evidence of chronology and palaeoecology of Sterkfontein, Swartkrans and Kromdraai from the fossil Bovidae. *Nature* **254**, 301–304 (1975).
157. Vrba, E. S. in *L'environnement des hominidés au Plio-Pléistocène* (eds Beden, M. et al.) 345–369 (Masson, 1985).
158. Churcher, C. S. & Watson, V. in *Swartkrans: A Cave's Chronicle of Early Man* (ed. Brain, C. K.) 137–150 (Transvaal Museum, Pretoria, 1993).
159. Blackwell, B. A. Problems associated with reworked teeth in electron spin resonance (ESR) dating. *Quat. Sci. Rev.* **13**, 651–660 (1994).
160. Steininger, C. Local ecological profile for *Paranthropus robustus* in South Africa using stable carbon isotopes from associated bovid teeth. *Quat. Int.* **279–280**, 466 (2012).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



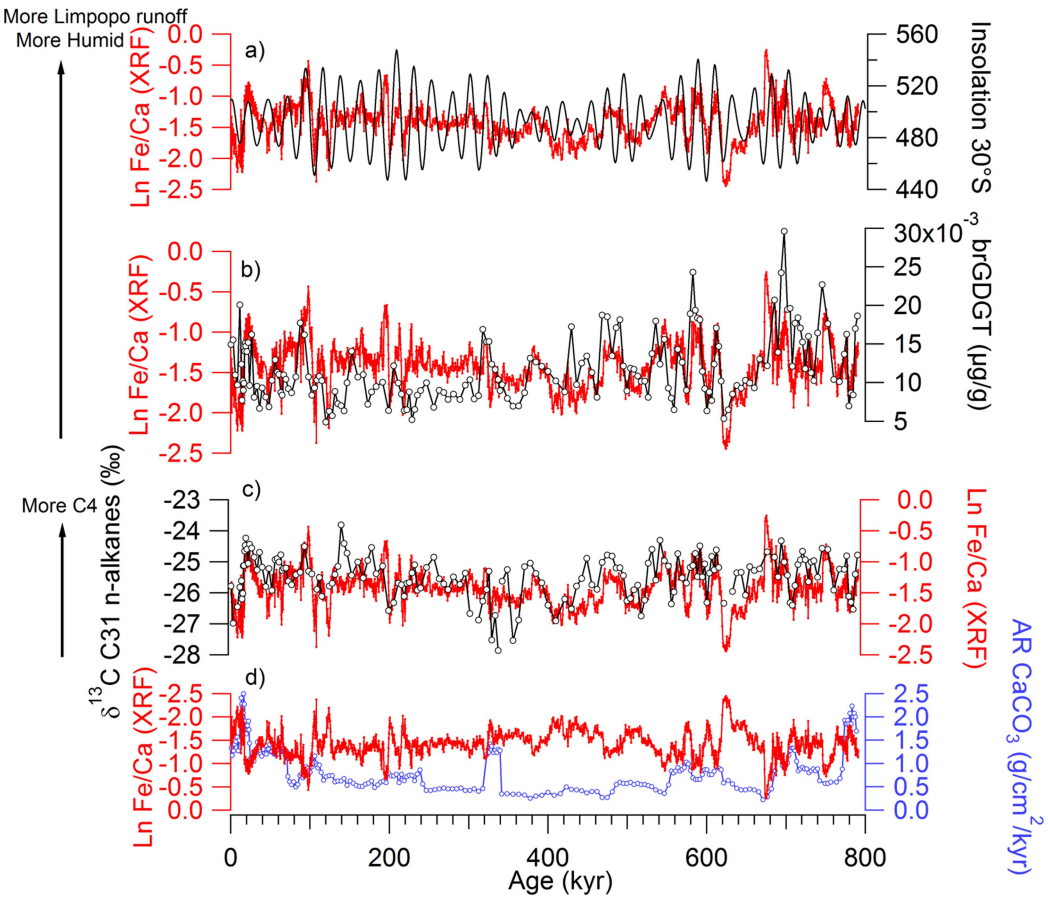
Extended Data Fig. 1 | ln(Fe/Ca) as a proxy for Limpopo runoff. Calcium and iron both have complex and multiple origins in marine sediments. Iron can be related to redox variations, detrital and fluvial input, among others, and calcium can be related to the biogenic fraction (foraminifera or nannofossils) and detrital input. To properly interpret the ln(Fe/Ca) ratio at our study location, we applied principal components analysis[120]. **a**, PC1 describes 66% of the total variance for the entire site MD96-2048. The negative loadings for PC1 are calcium and strontium, and all other elements (aluminium, silicon, potassium, titanium, iron and zirconium) have positive loadings. Calcium and strontium are associated with biogenic carbonate and are mainly related to presence of foraminifera. Element matrix correlation shows a strong positive linear correlation ($R > 0.70$) between iron and typically detrital elements, such as aluminium, silicon, titanium and potassium. Calcium shows negative correlation with iron ($R = -0.5$). **b**, ln(Fe/Ca) shows a strong correlation with PC1 ($R = 0.94$) and a strong relationship with Limpopo runoff proxies (Extended Data Fig. 3). Iron and titanium elements are related to terrigenous and siliciclastic components (heavy minerals and oxides) and the variation in carbonate content (calcium) is mainly due to dilution by terrigenous sediment. ln(Fe/Ca) is therefore a proxy of Limpopo runoff, consistent with previous studies in riverine basins throughout the African continent[10,121–124]. To confirm a weak influence of sea-level changes on the Fe/Ca record, we compared our ln(Fe/Ca) record with a previous reconstruction of the deep-water $\delta^{18}O$ component for relative sea level[125], (**b**, bottom). Both records are plotted against the LR04 chronology. Visual inspection and statistical testing do not support a dominant effect of sea-level changes on the ln(Fe/Ca) record ($R = 0.05$). PC3, which describes 11% of the total variance for the entire site MD96-2048, is closely related to sea-level changes. The negative loadings for PC3 are mainly strontium and, to a lesser degree, potassium and titanium, and the main positive loadings are zirconium and, to a lesser degree, silicon.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



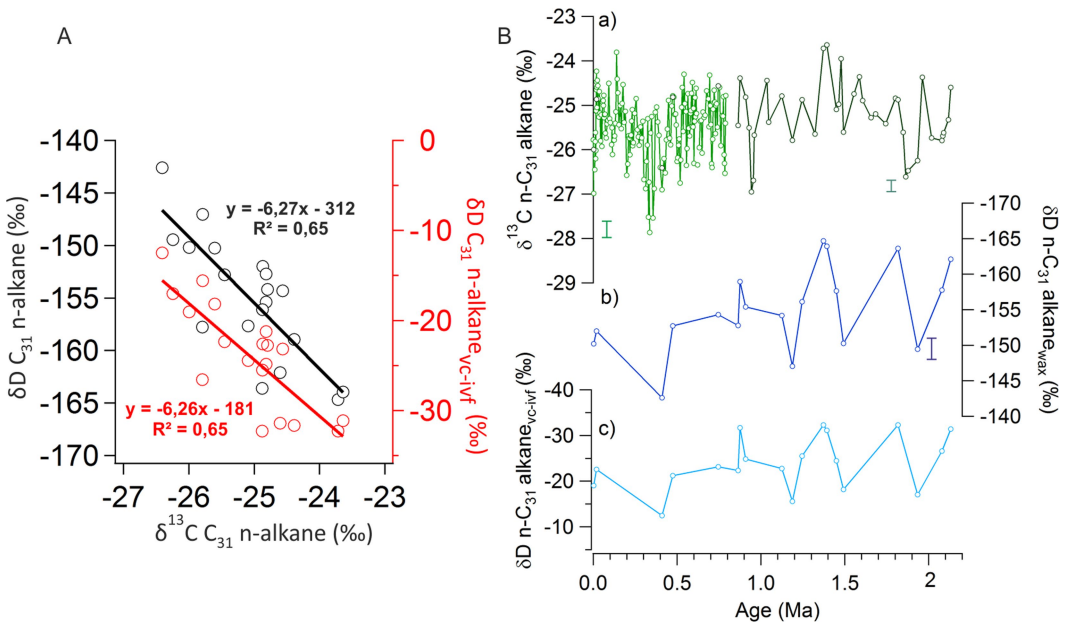
Extended Data Fig. 2 | Control on the δD composition of precipitation in the Limpopo catchment. a, **b**, Seasonal δD composition of precipitation (**a**) and amount of precipitation at Pretoria station[126] (**b**), in comparison to the results of the iLOVECLIM model at the corresponding latitude and longitude[48,49]. All data are centred on their annual average. Depleted δD values are indicative of increasing amounts of rainfall[127]. **c**, Results of the transient simulation with the isotope-enabled numerical climate model iLOVECLIM for the δD composition of precipitation and precipitation in the Limpopo catchment (about −27.5° S to −22° S and 30° E to 36° E), for the past 150 kyr (Methods)[51]. Black curves show the results after filtering with a low-pass filter. The δD composition of precipitation and precipitation amount in the Limpopo catchment are negatively correlated ($R = -0.63$, $P \ll 0.001$) for the past 150 kyr. Maxima of precipitation are phased with maxima in austral summer insolation at 30° S and lead to more-depleted $\delta D_{precipitation}$ (amount effect).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



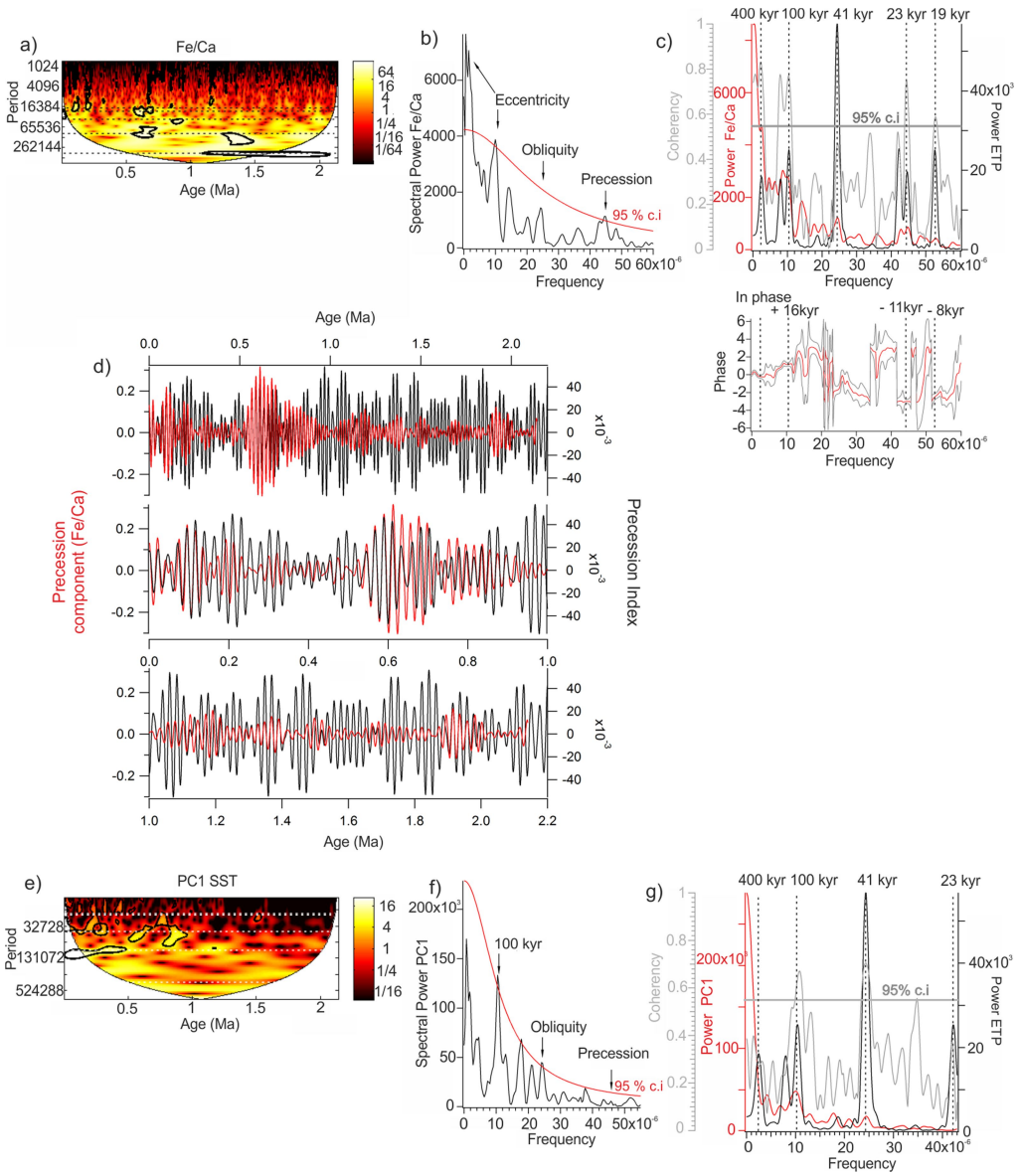
Extended Data Fig. 3 | Relationship between Limpopo runoff, local Southern Hemisphere insolation and the C_{31} *n*-alkane $\delta^{13}C$ record for the past 800 kyr. a, Comparison between the ln(Fe/Ca) XRF signal and austral summer local insolation at 30° S[31]. **b**, Comparison between the ln(Fe/Ca) XRF signal and the brGDGT concentration in the sediment[15]. brGDGTs are commonly found in soil and can be attributed to Limpopo River runoff[15]. **c**, Comparison between the ln(Fe/Ca) XRF signal and the C_{31} *n*-alkane $\delta^{13}C$ record[16]. An increased amount of Limpopo River discharge is associated with more C_4 plant input and an increase in austral summer insolation at 30° S. **d**, Comparison between inverted ln(Fe/Ca) XRF signal and the accumulation rate (AR) of $CaCO_3$ as a measure of biogenic carbonate. The ln(Fe/Ca) XRF record is not primarily controlled by dilution due to biological productivity ($R = 0.1$). A previous study of the past 0.8 Myr of core MD96-2048 interpreted shifts towards more-depleted $\delta^{13}C_{wax}$ as potentially reflecting more-humid conditions[16]. However, the anti-correlation between $\delta^{13}C_{wax}$ and δD_{wax} values (Extended Data Fig. 4) in our study indicates that enriched $\delta^{13}C_{wax}$ values are associated with more-humid conditions. Because C_4 plants in the Limpopo catchment are dominant in the interior (Fig. 1), we propose that more-enriched $\delta^{13}C_{wax}$ values indicate a higher relative contribution from sources located farther upstream (more C_4 plants) during times of high runoff, compared to only downstream sources (more C_3 plants) during times of low discharge. In addition, humid conditions would have favoured the extension of sedge-rich vegetation (Cyperaceae, of which 20–60% are C_4 plants in this region[128]) in riverine swamps and floodplains along the river course, explaining the detected increase in Cyperaceae pollen at times of increased fluvial discharge (Fig. 2). Studies of sediments from the adjacent Zambezi catchment similarly suggest the extension of swampy sedge-rich vegetation—including C_4-Cyperaceae—when river discharge was high, and infer that more C_4 plant waxes are exported to the ocean when the flooding of floodplains occurs during rainfall maxima[10,129].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



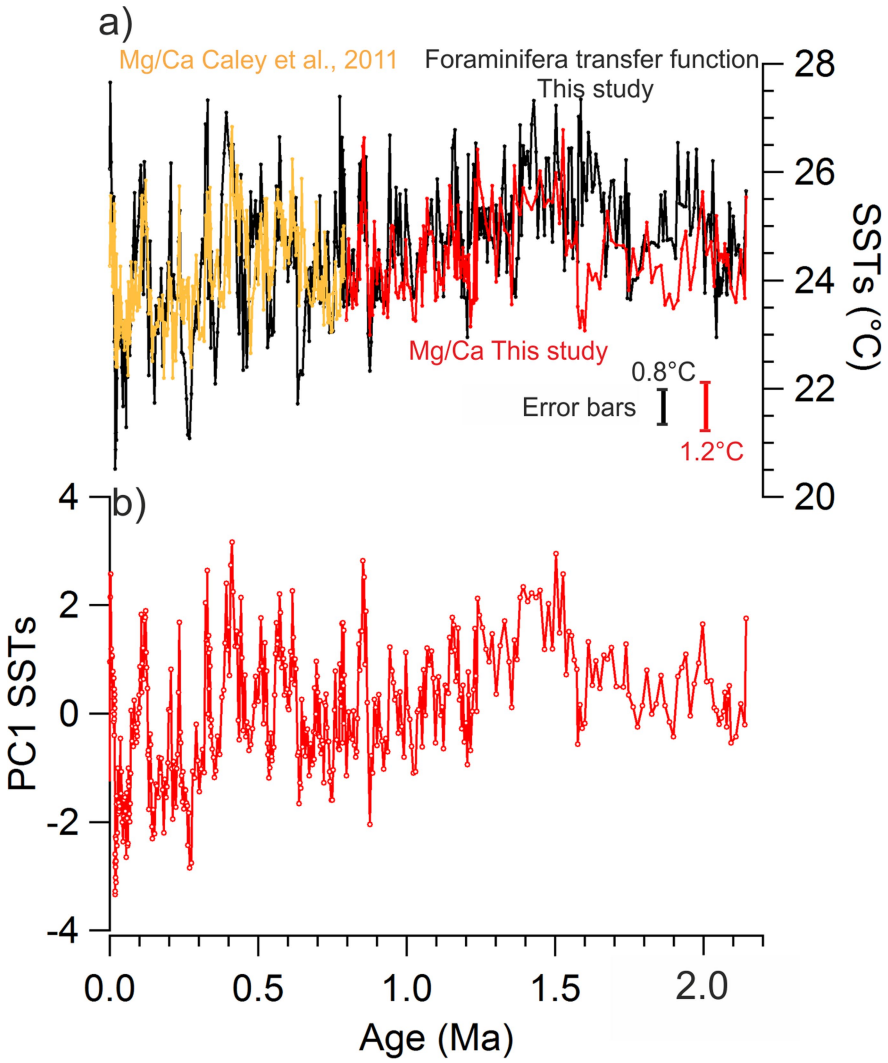
Extended Data Fig. 4 | Relation between the $\delta^{13}C$ C_{31} *n*-alkanes record and the δD C_{31} *n*-alkanes record. a, Correlation between the record of $\delta^{13}C$ C_{31} *n*-alkanes and the record of δD C_{31} *n*-alkanes, with or without vegetation and ice-volume correction (vc-ivf) over the past 2.14 Myr ($n = 19$ samples). An anti-correlation exists between the $\delta^{13}C$ and the δD signals of the C_{31} *n*-alkanes. The C_{31} *n*-alkane is used because it is the most abundant homologue in the samples. **b**, Raw $\delta^{13}C_{wax}$, δD_{wax} data and δD_{wax} adjusted for ice-volume and vegetation changes from core MD96-2048. Mean analytical uncertainties are indicated. Top, $\delta^{13}C_{wax}$ of the C_{31} homologue (data from a previous study[16] in light green, and data from this study in dark green). Middle, δD_{wax} of the C_{31} homologue. Bottom, δD_{wax} of the C_{31} homologue adjusted for ice-volume changes (ivf) using a seawater $\delta^{18}O$ curve[125] and converting to δD assuming an increase of 7.2‰ at the Last Glacial Maximum. We use 7.2‰ because measurements of sediment pore water $\delta^{18}O$ and δD suggest that the glacial ocean δD increase has a mean value of 7.2‰[130]. We also adjusted the δD_{wax} record for vegetation changes (vc) using published fractionation factors (−123‰ ± 31‰ for C_3 trees, −139‰ ± 27‰ for C_4 grasses[131]) and the $\delta^{13}C_{wax}$ signal following a previously published procedure[132]. End-member $\delta^{13}C_{wax}$ values used for C_3 and C_4 vegetation were −36‰ and −21.5‰, respectively[133]. The error ranges for the vegetation fractionation factors are very large[131]. They derive from the compilation of a global dataset from individual plants, which is not comparable to an ecosystem fractionation in a specific catchment (such in the Limpopo) that will fractionate with a much smaller uncertainty. However, as we do not know the exact fractionation factor in the Limpopo catchment and regard the uncertainties from the global compilation as unrealistic for a specific ecosystem we refrained from propagating this uncertainty into the vegetation corrections. The vegetation and ice-volume-adjusted δD_{wax} record is very similar to the unadjusted record, highlighting the fact that the adjustments have a minor effect.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



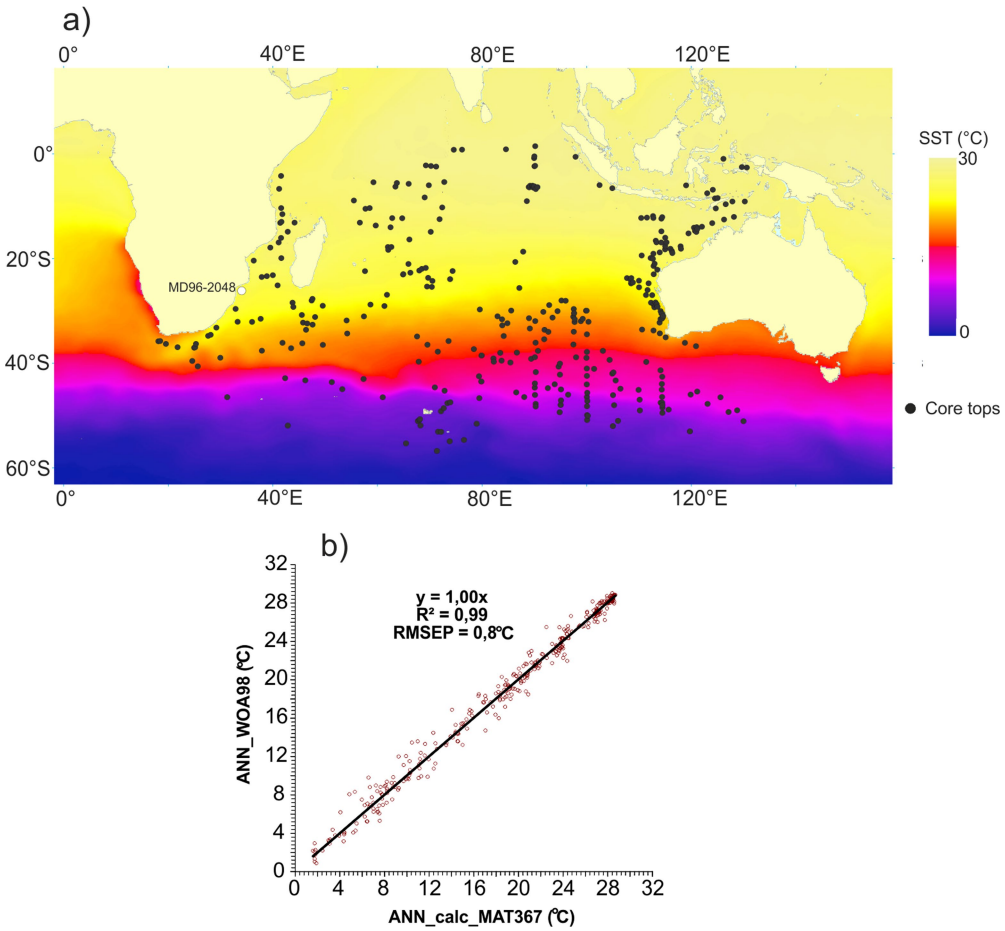
Extended Data Fig. 5 | Statistical analyses for the ln(Fe/Ca) XRF record and PC1 SST record. a, Spectral power for ln(Fe/Ca) by wavelet analysis realized with a previously published MatLab package[134]. The thick contour designates the 5% significance level against red noise. Dashed black lines indicate the variability at the precession, obliquity and eccentricity periods. **b**, Spectral analysis of ln(Fe/Ca) with REDFIT[135]. The red line shows the false-alarm level at the 95% confidence interval. Spectral peaks exceeding the false-alarm level can be considered significant[135]. **c**, Blackman–Tukey cross correlation between ln(Fe/Ca) XRF and eccentricity–tilt–precession (ETP) realized with the Analyseries software[37] for the past 2.14 Myr. ETP is constructed by normalizing and stacking eccentricity, tilt (obliquity) and negative precession to evaluate coherence and phase (timing) relative to orbital extremes[136]. The red curve shows the spectral power for ln(Fe/Ca) record. The black curve shows the spectral power for ETP. The coherency, which varies between 0 and 1, is represented by the grey curve and gives the interval within which the spectrum is significant. In our case, the non-zero coherency is higher than 0.55 and is significant at the 95% confidence interval (grey line). There are significant spectral peaks for eccentricity and precession but not for obliquity. The ln(Fe/Ca) XRF record and ETP are in phase at the 400-kyr period, the eccentricity leads by 16 kyr the ln(Fe/Ca) record at the 100-kyr period and the ln(Fe/Ca) record is in anti-phase with negative precession (in-phase with positive precession) at the 19- and 23-kyr periods. The three statistical analyses are consistent and indicate significant variability at the 400-, 100-, 23- and 19-kyr periods and insignificant variability at the 41-kyr period. **d**, Comparison between the precessional component of the ln(Fe/Ca) record (Gaussian filter frequency 1/23,000; bandwidth: 5×10^{-6}) obtained with the Analyseries software[37] and the precession index. Maxima of the ln(Fe/Ca) precession component are in phase with precession index maxima. The precession cycles in the ln(Fe/Ca) record appear particularly strong between about 0.9 and 0.6 Ma. **e**–**g**, The same statistical analyses as in **a**–**c**, respectively, but for the PC1 SST record. In **e**, dashed white lines indicate the variability at the precession, obliquity and eccentricity periods. The three statistical analyses indicate significant variability at the 100- and 41-kyr periods but not significant power for the 400-kyr and 23-kyr (precession) periods.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Reconstruction using SST proxies for core MD96-2048 for the past 2.14 Myr. a, Reconstruction of SST using two different methods: Mg/Ca reconstruction based on previous[15] and new data (Mg/Ca ratios were converted into temperature values by applying a previously established equation[40]) and foraminifera transfer function reconstruction using the modern analogue technique. Error bars represent the error on the calibrations[40] (Extended Data Fig. 7). **b**, Empirical orthogonal function analysis[47] of the two SST records for the past 2.14 Myr. PC1 contains 74% of the total variance for the past 2.14 Myr. Correlation between SST proxies and PC1 for the past 2.14 Myr is $R = 0.71$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Foraminifera transfer function used for core MD96-2048. a, Location of the modern database, composed of 367 core tops from the south Indian Ocean[45] with present-day SST from the World Ocean Atlas (WOA) 2009[29]. **b**, Test for the modern database[45] yielding to a precision of 0.8 °C for the annual SST reconstructions. Modern hydrological parameters were obtained from the WOA (1998) database using a previously developed tool (http://www.geo.uni-bremen.de/geomod/Sonst/Staff/csn/woasample.html).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Fossil finds, their location and associated ages

Site/area	Taxon	Stratigraphic unit	Methods	References	Estimated age of fossils
Malapa	*Australopithecus sediba*		Biochronology	Dirks et al. (137)	From 1.95 Ma to 1.78 Ma
			U-Pb dating of flowstones: basal flowstone 1 at 2.026 ± 0.021 Ma; capping flowtone 2 at 2.048 ± 0.140 Ma	Pickering et al. (138)	From 2.188 Ma to 1.908 Ma
			Paleomagnetism : reversed polarity of flowstone 2, normal polarity of fossiliferous sediments	Pickering et al. (138)	Older than 1.95 Ma
			Synthesis of flowstone dating and paleomagnetism, 1.977 ± 0.002 Ma	Pickering et al. (138)	**From 1.979 to 1.975 Ma**
Cooper's D	*Paranthropus robustus*		Biochronology	Berger et al. (139)	From 1.9 Ma to 1.6 Ma
			U-Pb dating of flowstones: basal flowstone CDD1 at 1.526 ± 0.088 Ma; younger flowstone CDD3, intercalated within fossiliferous sequence, at ca. 1.4 Ma (unprecise dating from 1.617 Ma to 1.413 Ma)	de Ruiter et al. (140)	**From 1.615 Ma to ca. 1.4 Ma (upper facies A and C) and younger than 1.4 Ma (lower facies A and C)**
Drimolen	*P. robustus*	Main Quarry Site	Biochronology	Keyser et al. (141)	From 2.0 Ma to 1.5 Ma
			Biochronology	Adams et al. (142)	**From 2.3 Ma to 1.6 Ma**
Kromdraai B	*P. robustus*		Paleomagnetism of flowstone above Mb. 3, reverse polarity, older than normal Olduvai event (between 1.95 Ma and 1.78 Ma)	Thackeray et al. (143)	Older than 1. 95 Ma
			Alternative interpretation of paleomagnetic data from Thackeray et al. (2002)	Herries et al. (144); Herries & Adams (145)	From 1.78 Ma to 1.65 Ma
			Biochronology (including hominins) and paleomagnetism	Braga et al. (146)	Older than 2.18 Ma
Gondolin	*P. robustus*	GD2*	Biochronology and paleomagnetism (Olduvai normal polarity event of fossiliferous sediments)	Herries et al. (147)	**Slightly older than 1.78 Ma**
	P. robustus	GD1*	Biochronology and paleomagnetism (end of Olduvai normal polarity event in basal flowstone)	Adams et al. (148)	**Slightly younger than 1.78 Ma**
Sterkfontein	*P. robustus*	Mb. 5B "Oldowan infill"	Biochronology and lithic typology	Kuman & Clarke (106)	From 2.0 Ma to 1.7 Ma
		Mb. 5B "Oldowan infill"	ESR on bovid teeth (seven dates): 1.328 ± 0.087 Ma; 1.315 ± 0.295 Ma; 1.185 ± 0.96 Ma; 1.265 ± 0.125 Ma; 1.620 ± 0.626 Ma; 0.965 ± 0.147 Ma; 1.24 ± 0.28 Ma; weighed mean = 1.32 ± 0.08 Ma	Curnoe (149) reinterpreted in Herries & Shaw (150)	From 1.40 Ma to 1.24 Ma
		Mb. 5B "Oldowan infill"	Synthesis of ESR, U-Pb dating, paleomagnetism	Herries & Shaws (150)	From ca. 1.4 Ma to 1.2 Ma
		Mb. 5B "Oldowan infill"	ESR on teeth from 0.965 ± 0.147 Ma to 1.328 ± 0.087 Ma ; weighted mean LU-ESR (excluding one tooth with large internal errors) = 1.223 ± 0.155 Ma	Curnoe (149) reinterpreted in Herries et al. (144)	from 1.415 Ma to 1.241 Ma (maximal); 1.112 Ma to 0.818 Ma (minimal); from 1.378 Ma to 1.223 (weighed mean)
		Mb. 5B "Oldowan infill"	Cosmogenic burial (^{26}Al/^{10}Be) dating of a quartz manuport 2.18 ± 0.21 Ma	Granger et al. (151)	From 2.39 Ma to 1.97 Ma
		Mb. 5B "Oldowan infill"	Synthesis of paleomagnetism, U-Pb dating, and ESR	Herries & Adams (145)	From 1.8 Ma to 1.5 Ma
		Mb. 5B "Oldowan infill"	Synthesis of biochronology and ESR	Herries et al. (144)	From 1.38 Ma to 1.07 Ma
Swartkrans	*P. robustus*	Mb. 1	Cosmogenic burial (^{26}Al/^{10}Be) dating on quartz 2.19 ± 0.08 Ma	Gibbon et al. (152)	From 2.27 Ma to 2.11 Ma
		Mb. 1	Cosmogenic burial (^{26}Al/^{10}Be) dating on quartz 1.80 ± 0.09 Ma	Gibbon et al. (152)	From 1.89 Ma to 1.71 Ma
		Mb. 1 Hanging Remnant	Synthesis of U-Pb dating and ESR	Herries & Adams (145)	From 2.0 Ma to 1.8 Ma
		Mb. 1 Hanging Remnant	ESR 2.02 ± 0.36 Ma; 2.07 ± 0.37 Ma; 1.68 ± 0.28 Ma; weighed mean = 1.96 Ma-1.70 Ma	Curnoe et al. (153) reinterpreted in Herries & Adams (145)	From 1.96 Ma to 1.70 Ma
		Mb. 1 Hanging Remnant	ESR LU 1.39 ± 0.18 Ma	Herries et al. (144)	From 1.57 Ma to 1.21 Ma
		Mb. 1 Hanging Remnant	ESR LU 1.92 ± 0.34 Ma	Herries et al. (144)	From 2.26 Ma to 1.58 Ma
		Mb. 1 Hanging Remnant	ESR LU 1.21 ± 0.22 Ma	Herries et al. (144)	From 1.43 Ma to 0.99 Ma
		Mb. 1 Lower Bank	U-Pb dating	Pickering et al. (154)	From 2.3 to 1.6 Ma
		Mb. 1	U-Pb dating on tooth 1.83 ± 1.38 Ma	Balter et al. (155) Vrba (156, 157);	From 3.21 to 0.45 Ma
		Mb. 1	Biochronology	Churcher & Watson (158)	ca. 1.7 Ma
		Mb. 1	U-Pb dating of basal flowstone 2.249 ± 0.077 Ma; top flowstone 1.706 ± 0.069 Ma ; closer to 2.0 Ma-1.8 Ma	Pickering et al. (154)	**Maximal range from 2.326 Ma to 1.637 Ma; minimal range from 2.172 Ma to 1.775 Ma**
Swartkrans	*P. robustus*	Mb. 2	U-Pb dating on tooth 1.36 ± 0.29 Ma	Balter et al. (155) Vrba (156, 157);	**From 1.65 Ma to 1.07 Ma**
		Mb. 2	Biochronology	Churcher & Watson (158)	ca. 1.5 Ma
		Mb. 2	Relative position to Mb. 1 dated by U-Pb	Pickering et al. (154)	Younger than ca. 1.7 Ma
Swartkrans	*P. robustus*	Mb. 3	Synthesis of U-Pb and ESR dating	Herries & Adams (145)	From 1.3 Ma to 0.6 Ma
		Mb. 3	Cosmogenic burial on quartz 0.96 ± 0.09 Ma	Gibbon et al. (152)	**From 1.05 Ma to 0.87 Ma**
		Mb. 3	ESR 0.65 ± 0.15 Ma	Cited in Herries and Adams (145)	From 0.8 Ma to 0.5 Ma
		Mb. 3	ESR 1.25 ± 0.09 Ma	Cited in Herries and Adams (145)	From 1.34 Ma to 1.16 Ma
		Mb. 3	U-Pb dating on tooth 0.83 ± 0.21 Ma	Balter et al. (155) Vrba (156, 157);	**From 1.04 Ma to 0.62 Ma**
		Mb. 3	Biochronology	Churcher & Watson (158)	ca. 1.0 Ma
		Mb. 3	ESR LU on two bovid teeth (four dates): 0.71 ± 0.90 Ma and 0.80 ± 015 Ma; 0.65 ± 0.15 Ma and 0.70 ± 0.11 Ma; mean = 0.72 ± 0.13 Ma	Blackwell (159)	From 0. 85 Ma to 0.59 Ma
		Mb. 3	Synthesis of biochronology and U-Pb	Herries et al. (144)	**From 1.04 Ma to 0.62 Ma**

Data in this table were taken from refs[106,137–160]. We consider the ages in bold as the best estimate. The different dating methods did not yield any agreement regarding the age of Kromdraai B and Sterkfontein member 5 'Oldowan Infill'. Therefore, no estimate is highlighted in bold and the stratigraphic ranges are not shown in Fig. 3. We favour U–Pb dates and cosmogenic burial of quartz dates rather than biochronology or electron spin resonance, although dates produced using the latter two methods are generally not inconsistent with those produced using the other methods[145]. **P. robustus* fossils were not found at Gondolin GD 1 and GD 2 but nearby ex situ. Given the close age between Gondolin GD 1 and GD 2 deposits and the limited extent of outcrops, it has been suggested[145] that the ex situ hominin specimens from Gondolin are dated to around 1.78 Ma.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | $\delta^{13}C$ enamel of hominin and contemporaneous herbivores and associated statistical parameters for different sites in the Limpopo catchment

		$\delta^{13}C$ (‰)							
Group	n	mean	median	mode(s) (2 ‰ interval)	SD	min	max	range	References
Cooper' D herbivores	45	-5,0	-4,7	int. 1: -12 to -10 ‰ ; int. 2: -2 to 0 ‰	3,9	-11,5	2,5	14,0	Steininger (101)
Gondolin GD2 herbivores	21	-2,9	1,0	int. 1: -10 to -8 ‰ ; int. 2: 0 to 2 ‰	5,3	-11,1	3,5	14,6	Adams (102)
Swartkrans Mb. 1 herbivores	56	-4,8	-3,8	int. 1: -10 to -8 ‰ ; int. 2: -4 to 0 ‰	4,2	-12,4	2,2	14,6	Lee-Thorp et al. (99); Sponheimer et al. (83); Steininger (160)
Swartkrans Mb. 1 *Paranthropus*	18	-7,2	-6,9	int.: -8 to -6 ‰	1,2	-9,6	-4,9	4,7	Lee-Thorp et al. (99); Sponheimer et al. (84, 83)
Swartkrans Mb. 1 *Homo*	3	-8,2	-8,2	int.: -10 to -8 ‰	0,9	-9,2	-7,1	2,1	Lee-Thorp et al. (100)
Swartkrans Mb. 2 herbivores	53	-4,8	-3,8	int. 1: -12 to -10 ‰ ; int. 2: -4 to -2 ‰	4,4	-12,9	2,2	15,1	Lee-Thorp et al. (99, 100); Steininger (160)
Swartkrans Mb. 2 *Paranthropus*	2	-9,1	-9,1			-10,0	-8,1	1,9	Lee-Thorp et al. (99)
Swartkrans Mb. 3 herbivores	12	-3,8	-2,2	int.: -4 to 0 ‰	3,3	-11,6	-0,5	11,1	Steininger (160)
Swartkrans Mb. 3 *Paranthropus*	1	-7,9	-7,9			-7,9	-7,9		Lee-Thorp et al. (99)

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0358-x

Globally rising soil heterotrophic respiration over recent decades

Ben Bond-Lamberty[1]*, Vanessa L. Bailey[2], Min Chen[1], Christopher M. Gough[3] & Rodrigo Vargas[4]

Global soils store at least twice as much carbon as Earth's atmosphere[1,2]. The global soil-to-atmosphere (or total soil respiration, R_S) carbon dioxide (CO_2) flux is increasing[3,4], but the degree to which climate change will stimulate carbon losses from soils as a result of heterotrophic respiration (R_H) remains highly uncertain[5–8]. Here we use an updated global soil respiration database[9] to show that the observed soil surface R_H:R_S ratio increased significantly, from 0.54 to 0.63, between 1990 and 2014 ($P = 0.009$). Three additional lines of evidence provide support for this finding. By analysing two separate global gross primary production datasets[10,11], we find that the ratios of both R_H and R_S to gross primary production have increased over time. Similarly, significant increases in R_H are observed against the longest available solar-induced chlorophyll fluorescence global dataset, as well as gross primary production computed by an ensemble of global land models. We also show that the ratio of night-time net ecosystem exchange to gross primary production is rising across the FLUXNET2015[12] dataset. All trends are robust to sampling variability in ecosystem type, disturbance, methodology, CO_2 fertilization effects and mean climate. Taken together, our findings provide observational evidence that global R_H is rising, probably in response to environmental changes, consistent with meta-analyses[13–16] and long-term experiments[17]. This suggests that climate-driven losses of soil carbon are currently occurring across many ecosystems, with a detectable and sustained trend emerging at the global scale.

The sensitivity of R_H to ongoing changes in temperature, precipitation and organic matter input to the soil system remains highly uncertain. Because of the large stocks of global soil organic carbon content (SOC) and quickly changing climatic conditions, this uncertainty has large implications for predicting future dynamics of Earth's climate system and carbon (C) cycle[5]. But because R_H observations are infrequently performed, necessarily small-scale and highly variable, it is difficult to detect and attribute annual-to-decadal changes at larger spatial scales.

One way to quantify and interpret shifts in the flux of C from SOC is to examine the changing ratios between R_H and other parts of the C cycle. For example, widespread increases over time in the ratio of R_H to the larger R_S flux, a ratio which typically varies in a predictable manner[18], would suggest that R_H is rising. Quantifying this change would provide new constraints on whether greater mineralization of SOC is occurring, and the degree to which increased inputs by enhanced gross primary production (GPP) are affecting R_H[8,19]. We used a global soil respiration database (SRDB)[9], expanded and updated to include studies reporting data through 2014, to test these possibilities.

The mean R_H:R_S ratio observed in SRDB has risen over time, from 0.54 ± 0.18 in 1990–1998 to 0.63 ± 0.16 in 2007–2014 (Fig. 1). This change is significant ($P = 0.009$, $n = 318$) when ecosystem mean annual temperature (MAT) and mean annual precipitation (MAP) are controlled for in a linear model (Table 1, Extended Data Table 1, Extended Data Fig. 1). This trend in the R_H:R_S ratio was not induced by a site selection bias, as this model also controls for the possibility of researchers sampling more disturbed sites, sites with varying land cover or SOC stocks, and using different methods to partition R_H from R_S. The SRDB covers most of the global climate space (from −13.2 °C to 27.9 °C MAT, and from 63 mm to 4,563 mm MAP; Extended Data Fig. 2); examining climate controls[20] on R_H, both MAP ($P < 0.001$, exerting a positive effect), and potential evapotranspiration ($P = 0.006$) had the strongest effects on annual R_H, with disturbance and temperature exerting complex interactive effects.

A rising R_H:R_S ratio could be due to rising SOC losses and thus a climate feedback, and/or increasing GPP rates enhancing detritus inputs and thus counterbalancing C losses from SOC. To distinguish these possibilities, we examined the ratio of all soil-derived respiration fluxes to GPP[21], the ultimate source of both the autotrophic and heterotrophic soil surface CO_2 fluxes. We note that the more relevant ratio to study is that of R_H to net primary production (NPP) rather than to GPP,

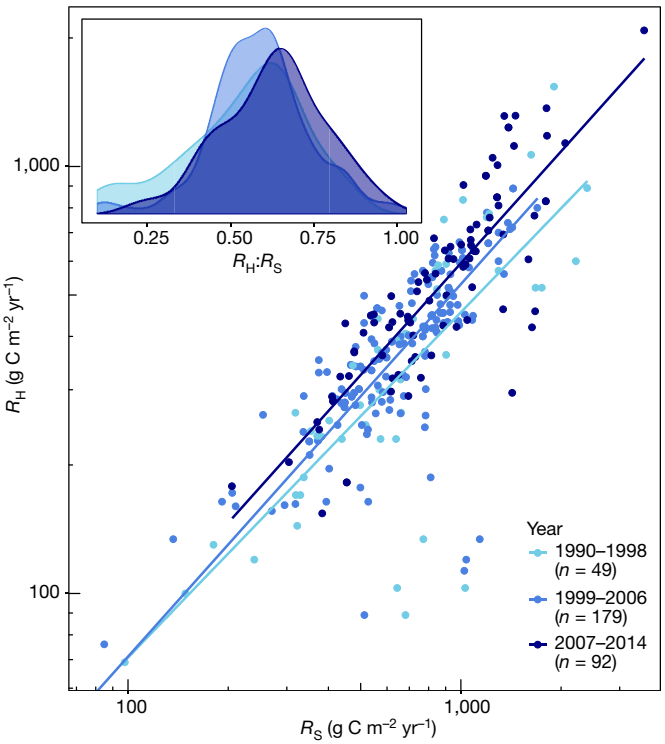


Fig. 1 | Changes in R_H:R_S over time. Data are drawn from unmanaged, unmanipulated ecosystems in a global soil respiration database[9], binned into 8–9 year groups (this binning is for display only, and was not part of the statistical analysis summarized in Table 1), and shown with linear regression lines; for clarity, one extremely low value is not shown. Note logarithmic axes. Inset, changes in R_H:R_S density distribution over time. See also Extended Data Table 1.

[1]Pacific Northwest National Laboratory, Joint Global Change Research Institute at the University of Maryland–College Park, College Park, MD, USA. [2]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA. [3]Department of Biology, Virginia Commonwealth University, Richmond, VA, USA. [4]Department of Plant and Soil Sciences, University of Delaware, Newark, DE, USA. *e-mail: bondlamberty@pnnl.gov

© 2018 Springer Nature Limited. All rights reserved.

Table 1 | Effects in linear model of how R_H:R_S changes over time

Effect	Degrees of freedom	Sum of squares	Mean sum of squares	*F*	*P*
Year	1	0.151	0.159	6.972	0.009
Disturbance	1	0.018	0.018	0.826	0.364
Partitioning method	10	0.363	0.036	1.677	0.086
MAT	1	0.439	0.439	20.293	<0.001
MAP	1	0.005	0.005	0.238	0.626
SOC	1	0.408	0.408	18.877	<0.001
Year * Method	7	0.432	0.062	2.853	0.007
MAT * MAP	1	0.075	0.075	3.460	0.064
Year * SOC	1	0.058	0.058	2.656	0.104
Residuals	293	6.339	0.022		

This table summarizes the statistically significant effects in the linear model examining how R_H:R_S changes over time. Effects tested include year of observation, disturbance (ecosystem coded as aggrading versus mature), method used to partition R_H from R_S, land cover (deciduous forest, evergreen forest, grassland, savannah, other), MAT, MAP and SOC. Terms that were not significant in the final model do not appear in the table. An asterisk denotes an effect interaction.

but global NPP estimates are subject to more uncertain assumptions. If increased SOC mineralization rates are raising R_H, one would expect both R_H and the total surface efflux R_S to rise relative to GPP.

The available data support this expectation as both R_H (the variable of primary interest, but for which far fewer measurements are available) and R_S (which incorporates both R_H and belowground autotrophic respiration) exhibit significant positive trends over time ($P = 0.012$ for R_H:GPP_{MTE}, $P < 0.01$ for all others; Fig. 2a) relative to GPP. These results are controlled, as above, for climate, land cover, partitioning method and disturbance. Grasslands exhibited much stronger R_S:GPP trends than forests for both the MODIS[11] and MTE[10] GPP datasets, perhaps because of their frequently high belowground carbon allocation.

Solar-induced chlorophyll fluorescence (SIF) is correlated with GPP and provides an independent test of how R_H is related to carbon assimilation over time. We examined the ratios of R_H and R_S to two global SIF data products (SCIAMACHY and GOME-2; see Methods). Both R_H and R_S exhibited temporal trends against the SCIAMACHY data after controlling for disturbance and measurement method ($P < 0.001$ for both; Fig. 2a). No R_H:SIF temporal trend was found using the GOME-2 SIF data, and R_S:SIF was only marginally significant ($P = 0.045$), but this is by far the shortest (2007–2014) of the datasets tested.

It is important to note that remotely sensed GPP and SIF estimates may not fully capture GPP increases owing to, for example, CO_2 fertilization[22]. The results above are, however, robust to such potential 'missed' GPP, even if one assumes a high rate of CO_2-driven GPP increase completely missed by satellite-based products (Extended Data Figs. 4 and 5). We also found a significant ($P = 0.008$, $n = 263$) temporal R_H:GPP trend using GPP from a wide range of global models[23] that include CO_2 fertilization mechanisms (Fig. 2b), again controlling for the previously mentioned variables. This underscores the robustness of the trends shown in Fig. 2a.

A second challenge to the above analysis is the large spatial mismatch between remotely sensed GPP and SIF data (over 10^9 m^2) versus R_H and R_S measurements (about 1 m^2). To address this, we used eddy-covariance C flux data reported in the FLUXNET2015 database to examine temporal changes in co-located ecosystem respiration and production. Nighttime net ecosystem exchange (NEE_{night}) is generally dominated[24] by R_S. Consequently, NEE_{night} might increase relative to GPP over time, if R_H and thus R_S are rising. We used the full 'Tier 1' dataset ($n = 1{,}162$) to test this expectation and found that the annual NEE_{night}:$GPP_{fluxnet}$ ratio is significantly rising with time ($P = 0.002$). We also attempted to use FLUXNET data to identify GPP measurements made in the same study site and year as site-specific R_S measurements, but from a bootstrap analysis, concluded that the small size of this dataset was unlikely to be adequate for this purpose (Extended Data Figs. 7 and 8).

A third problem is that, as in many previous global syntheses, the above analyses combine time and space responses. For a final test,

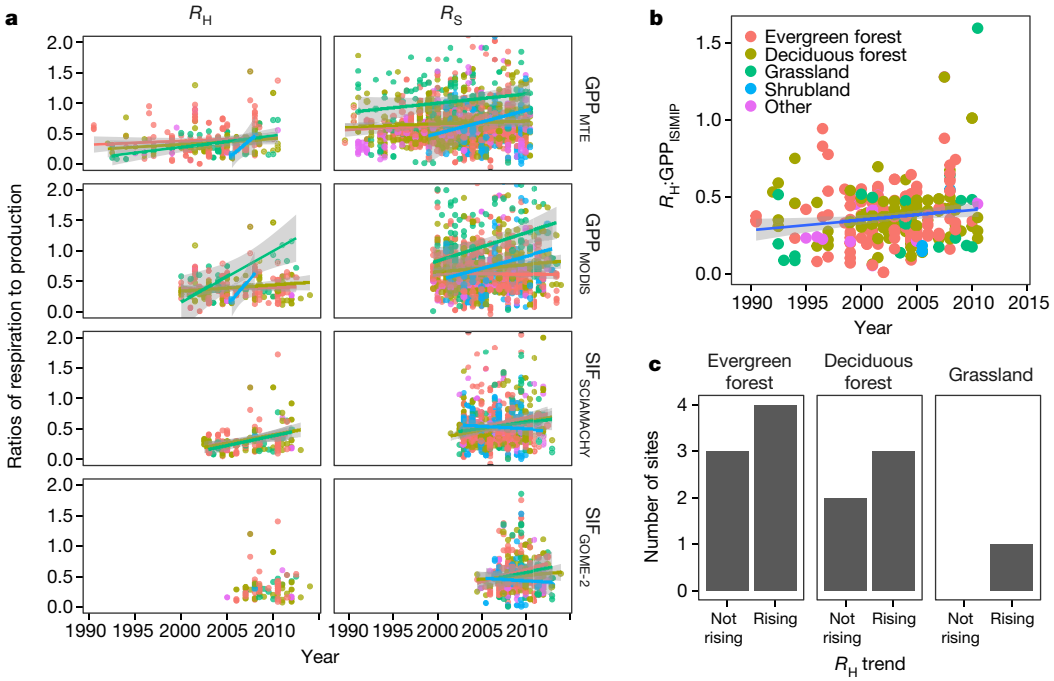


Fig. 2 | Changes in ratios of respiration to production over time. **a**, Changes in the ratio of respiration to GPP and SIF over time. Two respiration fluxes (R_H and R_S), two GPP sources (the MTE and MODIS datasets), and two SIF sources (the SCIAMACHY and GOME-2 datasets) are shown. For clarity, several high-ratio points are cut off. Points and linear regression lines are coloured by land cover (see key in **b**); lines in a panel imply that the overall temporal trend was significant in that panel. Grey shading shows 95% confidence intervals; residuals from these models are shown in Extended Data Fig. 3. **b**, Changes in the ratio of field-measured R_H to GPP modelled by suite of land models[23] over time. The trend line shows the statistically significant ($P < 0.001$) positive temporal trend in R_H:GPP using the GPP of ISIMIP models (see Table 1) and accounting for climate, land cover, disturbance and so on. **c**, Site-specific R_H trends in managed, unmanaged and natural ecosystems. Linear trend (not rising, slope ≤ 0, versus rising, slope >0) of R_H for sites in the SRDB[9] reporting at least three annual R_H measurements over at least eight years.

© 2018 Springer Nature Limited. All rights reserved.

Table 2 | Evidence supporting a global rise in soil R_H

Test	*n*	Temporal trend?	Land cover significant?
Increasing R_H:R_S	318	Yes	No
Increasing R_H:GPP_{MTE}	266	Yes	No
Increasing R_S:GPP_{MTE}	1,622	Yes*	Yes
Increasing R_H:GPP_{MODIS}	251	Yes	Yes
Increasing R_S:GPP_{MODIS}	1,256	Yes	Yes
Increasing R_H:SIF_{SCIA}	183	Yes	No
Increasing R_S:SIF_{SCIA}	879	Yes	No
Increasing R_H:SIF_{GOME-2}	90	No	No
Increasing R_S:SIF_{GOME-2}	402	Yes	No
Increasing R_S:$GPP_{fluxnet}$	121	No	Yes
Increasing NEE_{night}:$GPP_{fluxnet}$	1,162	Yes	Yes
Increasing R_H:GPP_{ISIMIP}	263	Yes	No
R_H climate response	332	N/A	No

Tests examined changes in ratios of R_H to R_S; R_H and R_S to GPP and SIF; NEE_{night} to FLUXNET[12] GPP; ISIMIP[23] GPP; and the response of R_H to climate. Columns include number of observations (*n*), significance ($P < 0.05$) of a 1989–2014 temporal trend after climate and other factors are accounted for in a linear regression, and whether the temporal trend was influenced by land cover (coniferous forest, deciduous forest, grassland, savannah, other). All temporal trends were also significant in a Theil–Sen robust regression unless otherwise noted with an asterisk.

we examined longitudinal, site-level R_H records from both managed and unmanaged ecosystems. There are only 13 sites in SRDB with such long-term records, and 8 of these exhibit rising R_H (Fig. 2c); when these data are pooled and controlled for climate, mean R_H also exhibits a significant ($P < 0.001$) rising trend. This sample is very small and lies in a tighter climate change space than the main dataset (Extended Data Fig. 9), but the prevailing positive trend for a majority of sites is consistent with the analyses above given the site-specific diversity of factors controlling R_H responses to climate and atmospheric CO_2 changes. Several sensitivity analyses (see Methods) suggest that, because of their high variability and short record length, individual sites in both FLUXNET and SRDB remain unable to reliably detect an increase in R_H; the signal is only slowly emerging from the noise, even for sites with multi-decadal records and across ecosystem types (Extended Data Fig. 6).

In summary, we have shown that multiple independent results (Table 2) converge to point towards a consistent finding of rising global R_H. This could be explained by at least two mechanisms, which are not mutually exclusive.

First, increased R_H might be temporarily fuelled by shifts in SOC forms and availability, and thus increased substrates for SOC mineralization, with little or no change in total SOC[8]. For example, eddy-covariance measurements from >10-yr FLUXNET sites indicate[25,26] that increases in GPP are outpacing concurrent rises in ecosystem respiration, resulting in greater ecosystem C uptake and detritus production, and thus greater bioavailable C for microbial metabolism. Highly productive grasslands and deciduous forests might be expected to respond more quickly to abiotic drivers[27], and more slowly decomposing evergreen litter may even inhibit microbial activities[28], together potentially explaining the trends shown in Fig. 2. This would be consistent with a 2015 inferred (as a residual, not measured) land sink of 1.9 ± 0.9 Pg C yr^{-1} (ref. [29]), but such a substrate-driven mechanism is not likely to be permanent, given the imbalance between soil C inputs and outputs (Fig. 2a). The C mass balance reported by FLUXNET sites exhibiting increased GPP and litter production[25,26] also suggests that new substrate inputs are not rising fast enough to keep up with increasing R_H.

A second and more tenable possibility is that rising R_H reflects enhanced SOC mineralization driven by climate changes, in particular rising global temperatures. If (as we calculate) global R_H has risen by approximately 1.2% over 25 years, against a mean air temperature change of +0.7 °C, this would be broadly consistent with warming experiment syntheses[13–16]. Assuming that rising R_H does in fact reflect SOC losses, this raises a consistency problem with estimated increases in the terrestrial C sink[29]. It also poses a methodological challenge: small changes in R_H and SOC are difficult and perhaps impossible (see Methods) to detect at individual sites, even those with a decade or more of data, but will add up to substantial losses at the global scale.

A number of limitations should be noted. The SRDB (as well as most Earth-observing networks such as FLUXNET) is dominated by data from Northern Hemisphere, upland sites, and there are relatively few data from high latitudes and tropics (about 14% of SRDB for each), relative to the large areas and carbon pools of these biomes. Some of the data are also not fully independent: for example, R_H is usually measured independently of R_S, but occasionally (3–5% of these data) it is estimated by subtraction of autotrophic respiration from R_S, introducing an autocorrelation. Finally, any observational analysis such as this infers causality, and thus it is necessary to maintain and expand long-term manipulative experiments[13,17].

These results pose new scientific challenges and opportunities for model benchmarking, hypothesis generation and testing, ecological forecasting and experiments. Rigorously testing for global SOC changes will require improvements in measuring, reporting and verification protocols, as it is not currently possible to derive a time-variant SOC map from repeated inventories except in 'Tier 3' countries[30]. More open global datasets of SOC stocks and fluxes[9], and more broadly distributed long-term ecological study sites with systematic and repeated R_S, R_H, and SOC measurements will help to test this finding of rising global R_H in the face of a seemingly stable or increasing terrestrial C sink.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0358-x.

Received: 30 March 2017; Accepted: 11 June 2018;
Published online 1 August 2018.

1. Köchy, M., Hiederer, R. & Freibauer, A. Global distribution of soil organic carbon—Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world. *Soil* **1**, 351–365 (2015).
2. Scharlemann, J. P. W., Tanner, E. V. J., Hiederer, R. & Kapos, V. Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Manag.* **5**, 81–91 (2014).
3. Bond-Lamberty, B. & Thomson, A. M. Temperature-associated increases in the global soil respiration record. *Nature* **464**, 579–582 (2010).
4. Hashimoto, S. et al. Global spatiotemporal distribution of soil respiration modeled using a global database. *Biogeosciences* **12**, 4121–4132 (2015).
5. Friedlingstein, P. et al. Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *J. Clim.* **27**, 511–526 (2014).
6. Zhou, T., Phi, P., Hui, D. & Luo, Y. Global pattern of temperature sensitivity of soil heterotrophic respiration (Q_{10}) and its implications for carbon-climate feedback. *J. Geophys. Res. Biogeosci.* **114**, G02016 (2009).
7. Trumbore, S. E. & Czimczik, C. I. An uncertain future for soil carbon. *Science* **321**, 1455–1456 (2008).
8. Giardina, C. P., Litton, C. M., Crow, S. E. & Asner, G. P. Warming-related increases in soil CO_2 efflux are explained by increased below-ground carbon flux. *Nat. Clim. Change* **4**, 822–827 (2014).
9. Bond-Lamberty, B. & Thomson, A. M. A global database of soil respiration data. *Biogeosciences* **7**, 1915–1926 (2010).
10. Jung, M. et al. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res. Biogeosci.* **116**, G00J07 (2011).
11. Zhao, M., Heinsch, F. A., Nemani, R. R. & Running, S. W. Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote Sens. Environ.* **95**, 164–176 (2005).
12. Baldocchi, D. D. 'Breathing' of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems. *Aust. J. Bot.* **56**, 1–26 (2008).
13. Crowther, T. W. et al. Quantifying global soil carbon losses in response to warming. *Nature* **540**, 104–108 (2016).
14. Lu, M. et al. Responses of ecosystem carbon cycle to experimental warming: a meta-analysis. *Ecology* **94**, 726–738 (2013).
15. Wang, X. et al. Soil respiration under climate warming: differential response of heterotrophic and autotrophic respiration. *Glob. Change Biol.* **20**, 3229–3237 (2014).
16. Zhou, L. et al. Interactive effects of global change factors on soil respiration and its components: a meta-analysis. *Glob. Change Biol.* **22**, 3157–3169 (2016).
17. Melillo, J. M. et al. Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. *Science* **358**, 101–105 (2017).

© 2018 Springer Nature Limited. All rights reserved.

18. Bond-Lamberty, B., Wang, C. & Gower, S. T. A global relationship between the heterotrophic and autotrophic components of soil respiration? *Glob. Change Biol.* **10**, 1756–1766 (2004).
19. Davidson, E. A. & Janssens, I. A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **440**, 165–173 (2006).
20. Hursh, A. et al. The sensitivity of soil respiration to soil temperature, moisture, and carbon supply at the global scale. *Glob. Change Biol.* **23**, 2090–2103 (2017).
21. Vargas, R. et al. On the multi-temporal correlation between photosynthesis and soil CO_2 efflux: reconciling lags and observations. *New Phytol.* **191**, 1006–1017 (2011).
22. Anav, A. et al. Spatiotemporal patterns of terrestrial gross primary production: a review. *Rev. Geophys.* **53**, 2015RG000483 (2015).
23. Warszawski, L. et al. The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): project framework. *Proc. Natl Acad. Sci. USA* **111**, 3228–3232 (2014).
24. Falge, E. et al. Seasonality of ecosystem respiration and gross primary production as derived from FLUXNET measurements. *Agric. For. Meteorol.* **113**, 53–74 (2002).
25. Pilegaard, K., Ibrom, A., Courtney, M. S., Hummelshøj, P. & Jensen, N. O. Increasing net CO_2 uptake by a Danish beech forest during the period from 1996 to 2009. *Agric. For. Meteorol.* **151**, 934–946 (2011).
26. Urbanski, S. P. et al. Factors controlling CO_2 exchange on timescales from hourly to decadal at Harvard Forest. *J. Geophys. Res.* **112**, G02020 (2007).
27. Keenan, T. F. et al. Increase in forest water-use efficiency as atmospheric carbon dioxide concentrations rise. *Nature* **499**, 324–327 (2013).
28. Adamczyk, S., Adamczyk, B., Kitunen, V. & Smolander, A. Monoterpenes and higher terpenes may inhibit enzyme activities in boreal forest soil. *Soil Biol. Biochem.* **87**, 59–66 (2015).
29. Le Quéré, C. et al. Global carbon budget 2016. *Earth Syst. Sci. Data* **8**, 605–649 (2016).
30. Vargas, R., Paz, F. & de Jong, B. Quantification of forest degradation and belowground carbon dynamics: ongoing challenges for monitoring, reporting and verification activities for REDD+. *Carbon Manag.* **4**, 579–582 (2013).

Acknowledgements We are indebted to the thousands of researchers who measured and published the data collected here. This work used eddy-covariance data acquired and shared by the FLUXNET community (see Methods). This research was supported by the US Department of Energy, Office of Science, Biological and Environmental Research as part of the Terrestrial Ecosystem Sciences Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. R.V. acknowledges support from NASA-CMS (80NSSC18K0173) and USDA (2014-67003-22070). C.M.G. received additional support from the National Science Foundation Division of Environmental Biology, Award 1353908.

Reviewer information *Nature* thanks A. Konings, K. Ogle and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.B.-L. conceived this study, and together with C.M.G. and R.V. designed the analysis. M.C. provided SIF data processing and expertise, and with V.L.B. furnished key insights. B.B.-L. wrote the manuscript in close collaboration with all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0358-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0358-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to B.B.-L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Datasets. This study used version v20170630a of the Global Soil Respiration Database (SRDB)[9], downloaded 30 June 2017 from https://github.com/bpbond/srdb. We used only records in the SRDB from studies that (i) reported annual R_H and/or R_S; (ii) had spatial (longitude and latitude) and temporal (measurement years) information; (iii) took place in natural or unmanaged ecosystems; (iv) had no experimental manipulation; and (v) used infrared gas analysers or gas chromatography (as opposed to for example, soda lime measurements that may underestimate CO_2 fluxes).

In general, the SRDB is oriented to seasonal to annual R_S data, and focuses on spatial breadth (a large number of sites around the world) rather than temporal detail (it does not attempt to capture continuous R_S measurements, for example). The annual fluxes used here occasionally come from continuous year-round measurements, but more often are computed by authors from interpolated manual measurements (mean measurement separation, 21 days; mean annual coverage, 89%). Such interpolation generally produces robust annual flux estimates[31,32]. Each site (that is, each unique longitude/latitude/ecosystem type combination) has on average 1.8–2.0 years of data, although a few have longer available records (see below).

These data ($n = 334$ for R_H and $n = 1{,}852$ for R_S) were then spatially and temporally matched with a variety of ancillary datasets using the R 'raster' package version 2.5.8. Air temperature, precipitation and potential evapotranspiration came from the CRUTEM4 (Climatic Research Unit, University of East Anglia) climate data[33] downloaded 5 January 2017 from http://www.cru.uea.ac.uk/data. Because precipitation has higher uncertainty than temperature, and is only indirectly linked with the soil moisture that directly affects soil respiratory metabolism, we also examined the ESA Soil Moisture Climate Change Initiative (CCI) 'Combined' dataset, Version 02.2, downloaded 6 June 2017 from http://data.ceda.ac.uk/neodc/esacci/soil_moisture/data/daily_files/COMBINED/v02.2/. Using these data instead of the CRU precipitation data did not change the overall results, but the ESA–CCI dataset has poor or no coverage in cloudy tropical regions (see ref. [34] for example and also http://www.esa-soilmoisture-cci.org/node/93), and when linked with the SRDB data under analysis here, 8% of data lacked soil moisture values, versus 5 (0.1%) lacking precipitation. We thus proceeded using precipitation only.

The bulk density and soil organic C content variables of the SoilGrids 1 km dataset[35] were downloaded on 8 January 2017 from https://soilgrids.com and used for a simple computation of 1 m SOC stock (as bulk density times C concentration). The geographic location of a few (around 10) soil respiration observations was missing in the SoilGrids dataset, and for these we substituted biome- and ecosystem-specific median values.

We used two global GPP datasets: the 1982–2011 MTE[10] GPP dataset, downloaded on 5 January 2017 from https://www.bgc-jena.mpg.de/bgi/index.php/Services/Overview, and the 2000–2015 MODIS GPP dataset[11] downloaded on 6 January 2017 from http://www.ntsg.umt.edu/project/modis/mod17.php. GPP data were used instead of NPP data because of the higher errors in the latter[10,36] due to uncertainties surrounding autotrophic respiration fluxes. We also extracted grid-cell-specific mean GPP from outputs of 8 global models in the ISIMIP[23] project, which uses community-agreed sets of scenarios with standardized climate variables and socio-economic projections.

For a more site-specific measure of GPP and C exchange, 'Tier 1' FLUXNET2015 data were downloaded on 30 January 2017 from http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/ and filtered for quality (NEE_VUT_REF_QC ≥ 0.5). FLUXNET GPP was linked to a given R_S and/or R_H measurement if both measurements occurred within 5 km, in the same ecosystem type, and in the same year.

This work used eddy-covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy-covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. Data collected by Beringer in the OzFlux network was funded under an ARC FT (FT1110602). Specific FLUXNET sites and years used can be found in the Supplementary Information.

Finally, we used the global Solar-Induced chlorophyll Fluorescence (SIF) dataset from the Global Ozone Monitoring Experiment-2 (GOME-2) instrument onboard the MetOp-A platform and the SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCIAMACHY). This chlorophyll fluorescence signal is strongly linked with GPP[37] (although the ratio of SIF to GPP varies by land cover, and thus the SIF analysis is unable to capture changes in GPP due to land cover change). The monthly mean quality-filtered and gridded GOME-2 and SCIAMACHY SIF data (that is, $SIF_{GOME\text{-}2}$ and $SIF_{SCIAMACHY}$) were used to reduce the high data noise. The gridded GOME-2 data[38] is available at 0.5° resolution from 2007 to 2015, and SCIAMACHY[39] is at 1° from 2003 to 2011. These two datasets are so far the longest global SIF records. Both are publicly available at ftp://ftp.gfz-potsdam.de/home/mefe/GlobFluo/. For convenience in visualizing temporal changes in R_H:GPP and R_H:SIF ratios, SIF data were rescaled to the range of GPP (0–3,500 g C m^2 yr^{-1}) in plots.

In general, our analysis assumed the following basic relationships[40] between various carbon cycle fluxes:

Total soil respiration is the sum of its (belowground) autotrophic and heterotrophic components:

$$R_S = R_{A(below)} + R_H$$

Autotrophic respiration is comprised of above- and belowground components:

$$R_A = R_{A(above)} + R_{A(below)}$$

NPP is the balance of GPP and autotrophic respiration:

$$NPP = GPP - R_A$$

In the absence of disturbance, NEE is equivalent to NPP minus R_H:

$$NEE \approx NPP - R_H$$

Statistical analysis. Some data were excluded a priori (that is, before constructing the analysis). These included: (i) studies performed before 1989, roughly when infrared gas analysers began to be widely used; (ii) several extremely high MODIS GPP values (>10 kg C m^{-2} yr^{-1}); (iii) 27 points with R_S:$GPP_{fluxnet} > 5$.

We used linear regression, weighted by years of R_H or R_S observation, to examine changes over time in the ratios of R_H:R_S, R_H:GPP, R_S:GPP, R_H:SIF and R_S:SIF, as well as the influence of climate on R_H. For each analysis detailed in the text, a full model was fitted for the variable of interest (for example, of R_H:R_S) that controlled for recent disturbance (the SRDB 'Stage' field, aggrading or mature), temperature and precipitation normals, and land cover; year of study had a first-order interaction with all these independent variables. Land-cover groupings included deciduous forests, $n = 782$; evergreen forests, $n = 1{,}058$; grasslands, $n = 270$; savannahs, $n = 116$; and other, $n = 321$. Both MAP and its square (that is, MAT^2) were included[20]. All models were examined for influential outliers and deviations from normality; no transformation of dependent variables was performed. Non-significant terms were then eliminated using a forward-and-back stepwise algorithm (using the R package 'MASS', version 7.3-47) based on the Akaike Information Criterion. Generally the text reports F statistics and P values from the analysis of variance (ANOVA) results of these linear models. A Theil–Sen estimator[41] was also computed for each temporal trend, independent of the linear regressions, using the 'mblm' R package version 0.12.

A linear model relating R_H to climate–annual air temperature (°C), precipitation (mm) and its square—to allow for a nonlinear response[20], potential evapotranspiration, recent disturbance, and leaf habit—was fitted (adjusted $R^2 = 0.25$, RSE = 23.3 g C m^{-2} yr^{-1}) and used to estimate global R_H. Global fluxes were based on 1989–2014 HadCRUT4[33] data and projected onto a 0.5° grid (grid area data downloaded from http://eos-webster.sr.unh.edu/).

About one-third of the SRDB R_S data have associated uncertainties (that is, an annual flux reported as $X \pm Y$) due to measurement error, spatial variability, and so on. The mean coefficient of variability for these R_S errors was 15%, and the median coefficient of variability was 10%. We did not attempt to propagate these uncertainties through our analysis, given their variable origin and incomplete nature, but a more complete treatment of uncertainty[42] will be important for testing the robustness of our results in the future.

Sensitivity analyses. We performed a number of tests to assess the robustness of our findings. To test whether the findings in Fig. 1 were an artefact of the 1989–2014 period chosen, the model summarized in Table 1 was repeatedly fitted to the SRDB dataset, as filtered with various first and final year assumptions (that is, the earliest and latest measurement allowed). We found that these results were not sensitive to choice of first and last years unless the timespan of the data was dramatically shortened (by at least 50%; Extended Data Fig. 1).

In addition, because satellites probably miss some large fraction of climate-driven terrestrial GPP increases[22], we examined the robustness of the results in Fig. 2a to this potential problem. First, we quantified the divergence between eddy covariance (FLUXNET) and remotely sensed GPP (Extended Data Fig. 4). Second, we asked how much GPP would satellites have to be missing to invalidate the GPP–respiration trends shown in Fig. 2a using a sensitivity analysis that assumed a very conservative (that is, high) GPP trend[43,44] of 0.5% yr^{-1} and repeatedly re-fitted the regressions shown in Fig. 2a assuming that satellites missed 0%, 10%, 20%, ..., 100% of this gain (Extended Data Fig. 5). Third, grid-cell-specific mean GPP was extracted from outputs of 15 global models in the ISIMIP[23,45] project—which

© 2018 Springer Nature Limited. All rights reserved.

uses community-agreed sets of scenarios with standardized climate variables and socio-economic projections, and whose GPP models include CO_2 fertilization—and compared to observed R_H (Fig. 2B).

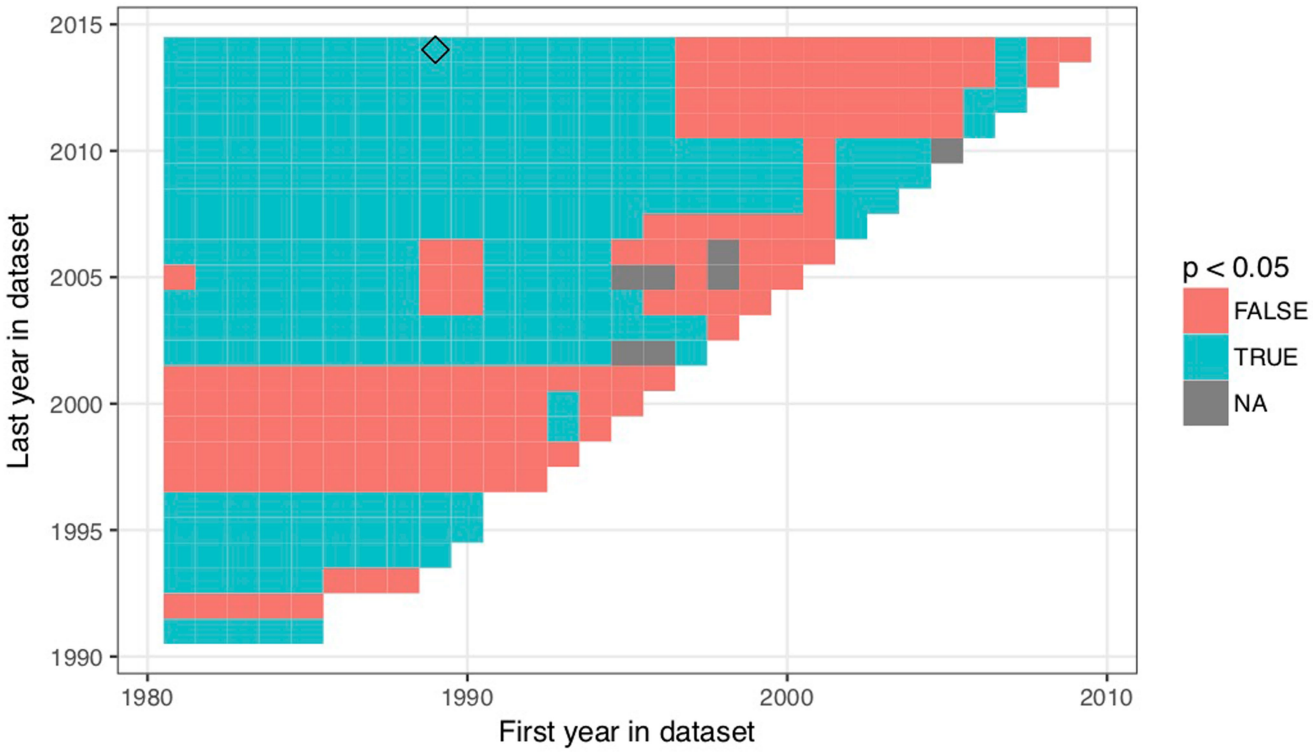
We also examined whether the inconsistency in longitudinal (site) results (Fig. 2c) fell within the range of what might naturally occur—that is, whether the variability was numerically consistent with the variability in global carbon flux datasets. To do so, we calculated the chance of observing a statistically significant R_H increase at some random location on Earth's surface over the last 25 years using a global gridded R_H dataset[4], by testing for a significant temporal trend in each of its 57,048 terrestrial grid cells. We calculated this probability to be 26%: in other words, there would be only a 1-in-4 chance of observing a significant R_H increase even with a quarter-century of perfect data ('perfect' in the sense of no random measurement error, as would occur in reality), because the climate-driven R_H increase is small compared to interannual R_H variation. The probability of observing a statistically significant R_H increase using only the last 10 years' data, comparable to the record length of most of the 13 longitudinal sites (Fig. 2c), was only 5%. A parallel exercise was performed with the global GPP datasets and yielded similar results. We also compared the sites' location in climate change space with that of the overall Fig. 1 dataset (Extended Data Fig. 9).

Finally, we performed a bootstrap analysis to quantify whether the FLUXNET data subset (that included GPP measurements made in the same study site and year as site-specific R_S measurements in the SRDB) was adequate, as this was a small dataset ($n = 106$ site-years, 19 sites) dominated by one site (Harvard Forest, with 56% of the available data). What was the likelihood that this dataset is simply too small to detect signals of rising R_H, given site-to-site variability in climate and carbon dynamics? To answer this we made different random draws from the original dataset, with 1,000 bootstrap draws per fraction of artificial no-trend data (Extended Data Figs. 7 and 8).

Code and data availability. The R language and environment for statistical computing[46] version 3.4.3 was used for all analyses. Code and data to reproduce all results are available at https://github.com/bpbond/rh-changes.

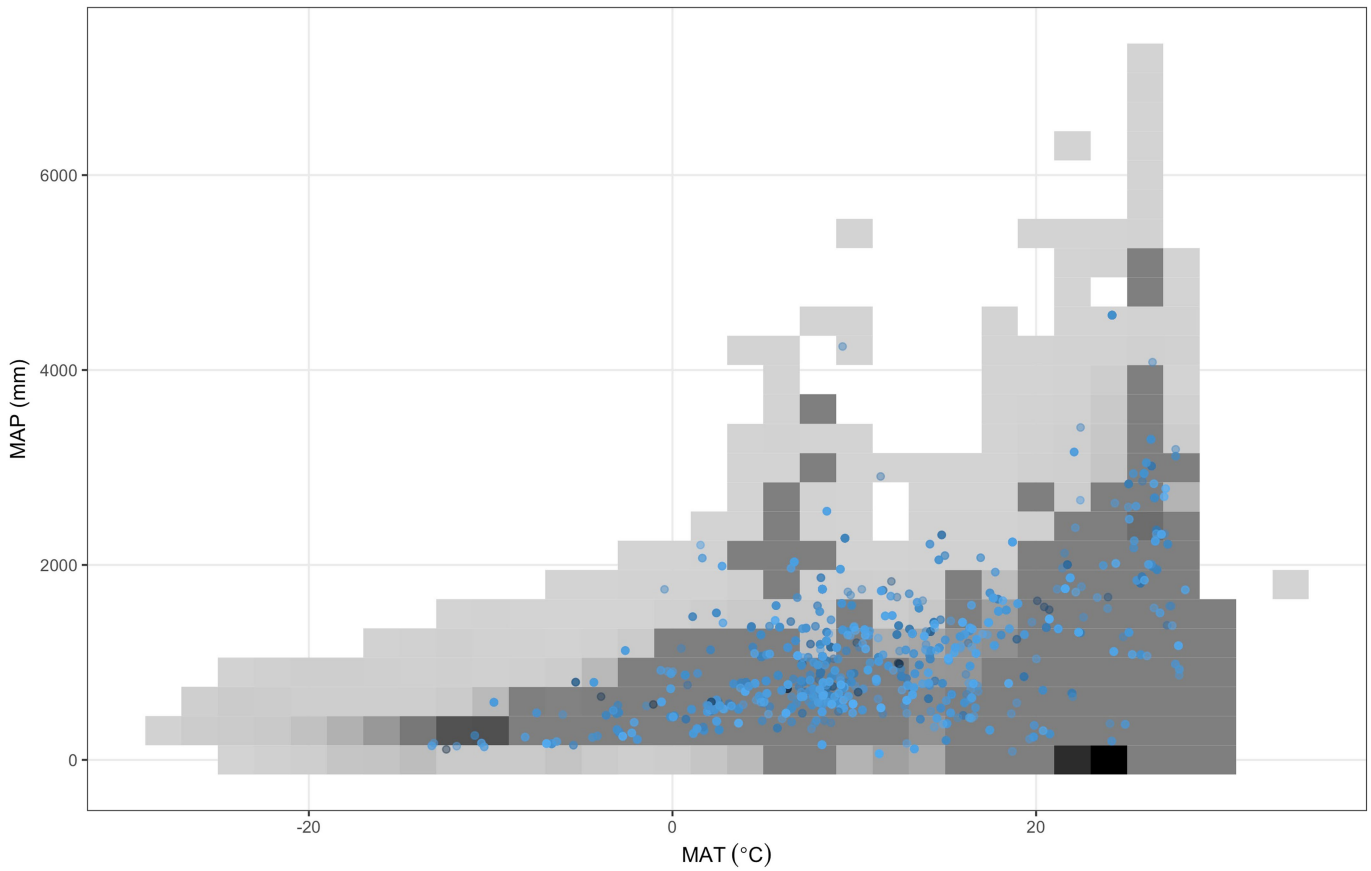
31. Wayson, C. A., Randolph, J. C., Hanson, P. J., Grimmond, C. S. B. & Schmid, H. P. Comparison of soil respiration methods in a mid-latitude deciduous forest. *Biogeochemistry* **80**, 173–189 (2006).
32. Reichstein, M. et al. Modeling temporal and large-scale spatial variability of soil respiration from soil water availability, temperature and vegetation productivity indices. *Glob. Biogeochem. Cycles* **17**, 1104 (2003).
33. Osborn, T. J. & Jones, P. D. The CRUTEM4 land-surface air temperature data set: construction, previous versions and dissemination via Google Earth. *Earth Syst. Sci. Data* **6**, 61–68 (2014).
34. Dorigo, W. et al. Evaluating global trends (1988–2010) in harmonized multi-satellite surface soil moisture. *Geophys. Res. Lett.* **39**, L18405 (2012).
35. Hengl, T. et al. SoilGrids1km — global soil information based on automated mapping. *PLoS ONE* **9**, e105992 (2014).
36. Ito, A. A historical meta-analysis of global terrestrial net primary productivity: Are estimates converging? *Glob. Change Biol.* **17**, 3161–3175 (2011).
37. Frankenberg, C. et al. New global observations of the terrestrial carbon cycle from GOSAT: patterns of plant fluorescence with gross primary productivity. *Geophys. Res. Lett.* **38**, L17706 (2011).
38. Joiner, J. et al. Global monitoring of terrestrial chlorophyll fluorescence from moderate-spectral-resolution near-infrared satellite measurements: methodology, simulations, and application to GOME-2. *Atmos. Meas. Tech.* **6**, 2803–2823 (2013).
39. Köhler, P., Guanter, L. & Joiner, J. A linear method for the retrieval of sun-induced chlorophyll fluorescence from GOME-2 and SCIAMACHY data. *Atmos. Meas. Tech.* **8**, 2589–2608 (2015).
40. Chapin, F. S. et al. Reconciling carbon-cycle concepts, terminology, and methods. *Ecosystems* **9**, 1041–1050 (2006).
41. Fernandes, R., Leblanc, S. G. Parametric (modified least squares) and non-parametric (Theil–Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. *Remote Sens. Environ.* **95**, 303–316 (2005).
42. Xu, M. & Shang, H. Contribution of soil respiration to the global carbon equation. *J. Plant Physiol.* **203**, 16–28 (2016).
43. Ito, A. et al. Photosynthetic productivity and its efficiencies in ISIMIP2a biome models: benchmarking for impact assessment studies. *Environ. Res. Lett.* **12**, 085001 (2017).
44. Kolby Smith, W. et al. Large divergence of satellite and Earth system model estimates of global terrestrial CO_2 fertilization. *Nat. Clim. Change* **6**, 306–310 (2016).
45. Rosenzweig, C. et al. Assessing inter-sectoral climate change risks: the role of ISIMIP. *Environ. Res. Lett.* **12**, 010301 (2017).
46. R Development Core Team. *R: A Language And Environment For Statistical Computing*. Version 3.3.3 (2017).

© 2018 Springer Nature Limited. All rights reserved.



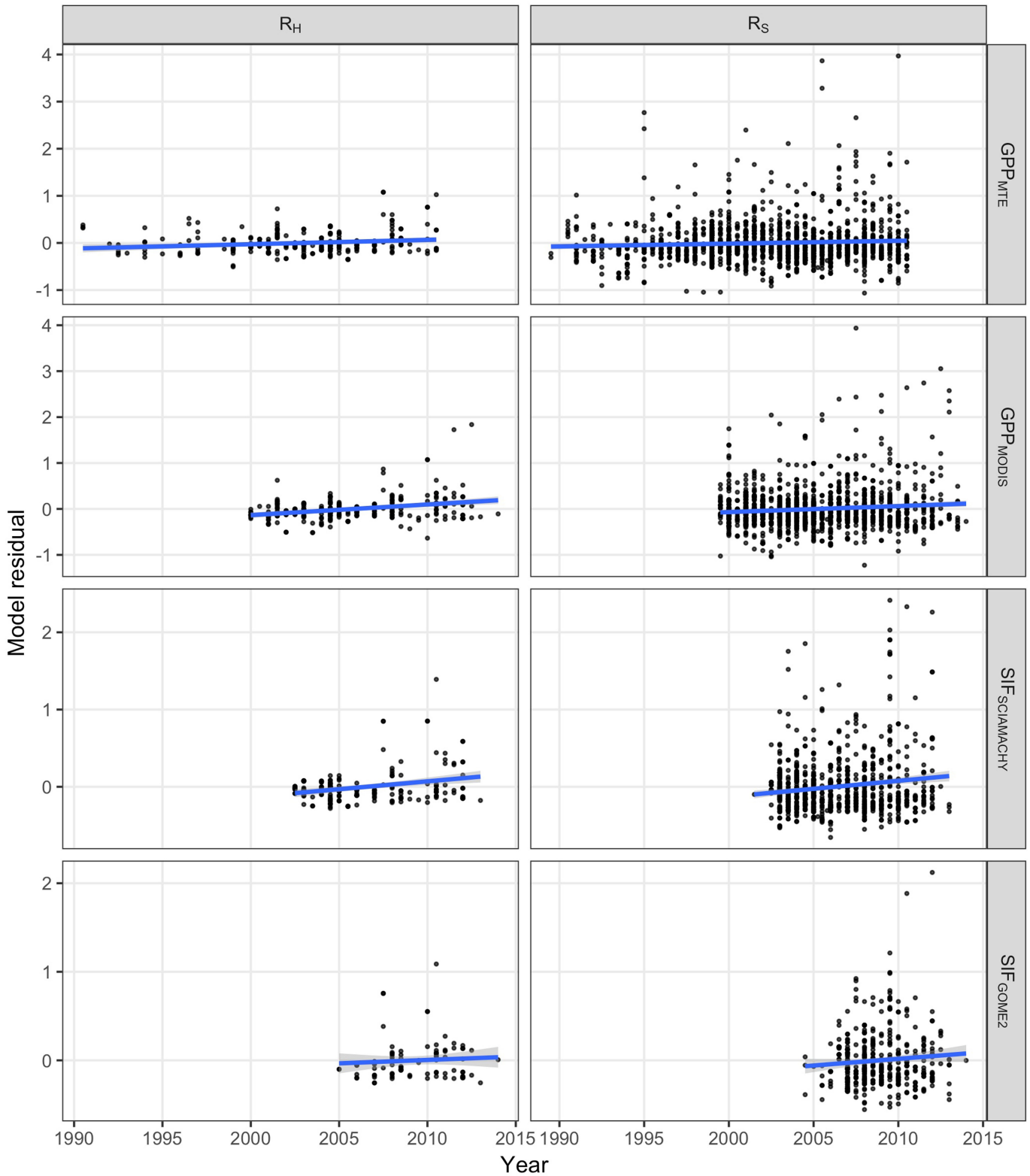
Extended Data Fig. 1 | Effect of dataset temporal span on R_H:R_S trend. The model summarized in Table 1 was repeatedly fitted to the SRDB dataset, as filtered with various first and final year assumptions (that is, the earliest and latest measurement allowed, to test whether the results were an artefact of the 1989–2014 period chosen, indicated by a black diamond in the figure). Colour shows the significance of the R_H:R_S temporal trend.

© 2018 Springer Nature Limited. All rights reserved.



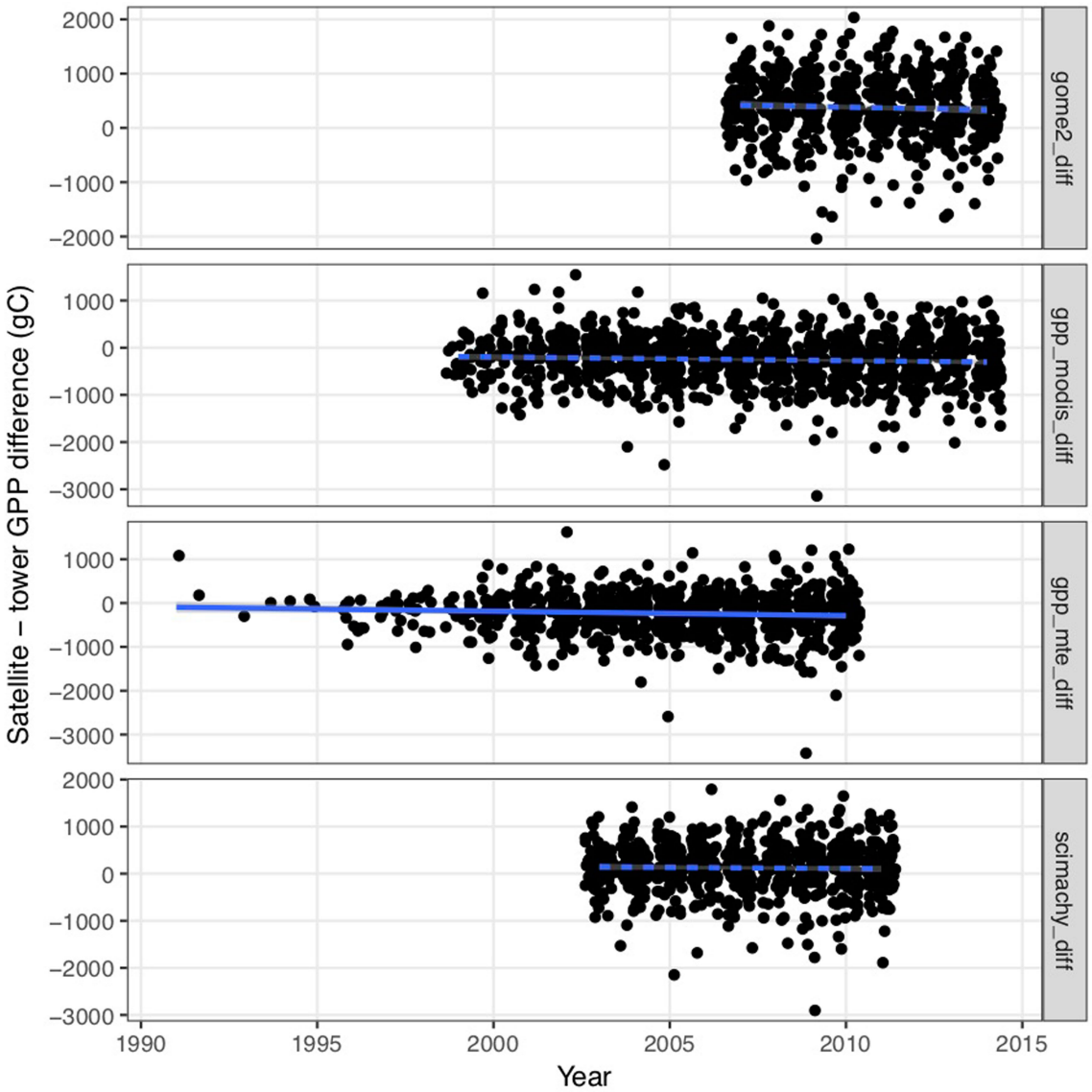
Extended Data Fig. 2 | SRDB coverage in terrestrial climate space. Blue dots show MAT and MAP of the entries in the database; lighter points indicate more recent data. Grey background indicates global climate distribution in the HadCRUT4[33] data, with darker shades indicating more land area, for 1991–2010.

© 2018 Springer Nature Limited. All rights reserved.



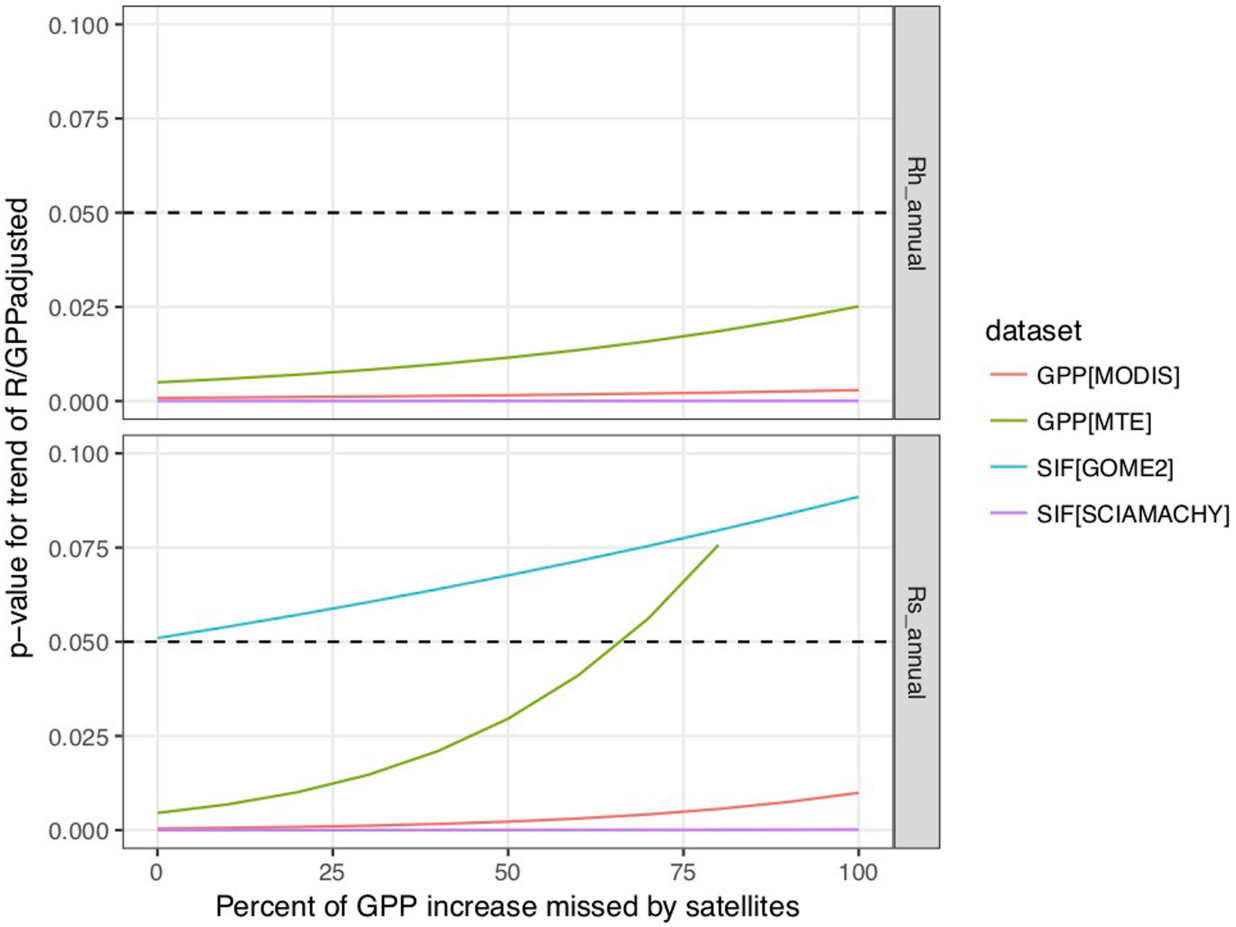
Extended Data Fig. 3 | Respiration-to-production model residuals over time. The model examines how the dependent variable—the ratio of respiration to GPP or SIF—is related to the independent variables: climate, land cover, disturbance and SOC content; for more details, see Methods. Two respiration fluxes (R_H and R_S), two GPP sources (the statistically upscaled MTE dataset, and the remotely sensed MODIS product), and two SIF sources (SCIAMACHY and GOME-2) are shown. Grey bands show 95% confidence intervals. Blue lines indicate least-squares trend, while grey bands show 95% confidence intervals.

© 2018 Springer Nature Limited. All rights reserved.



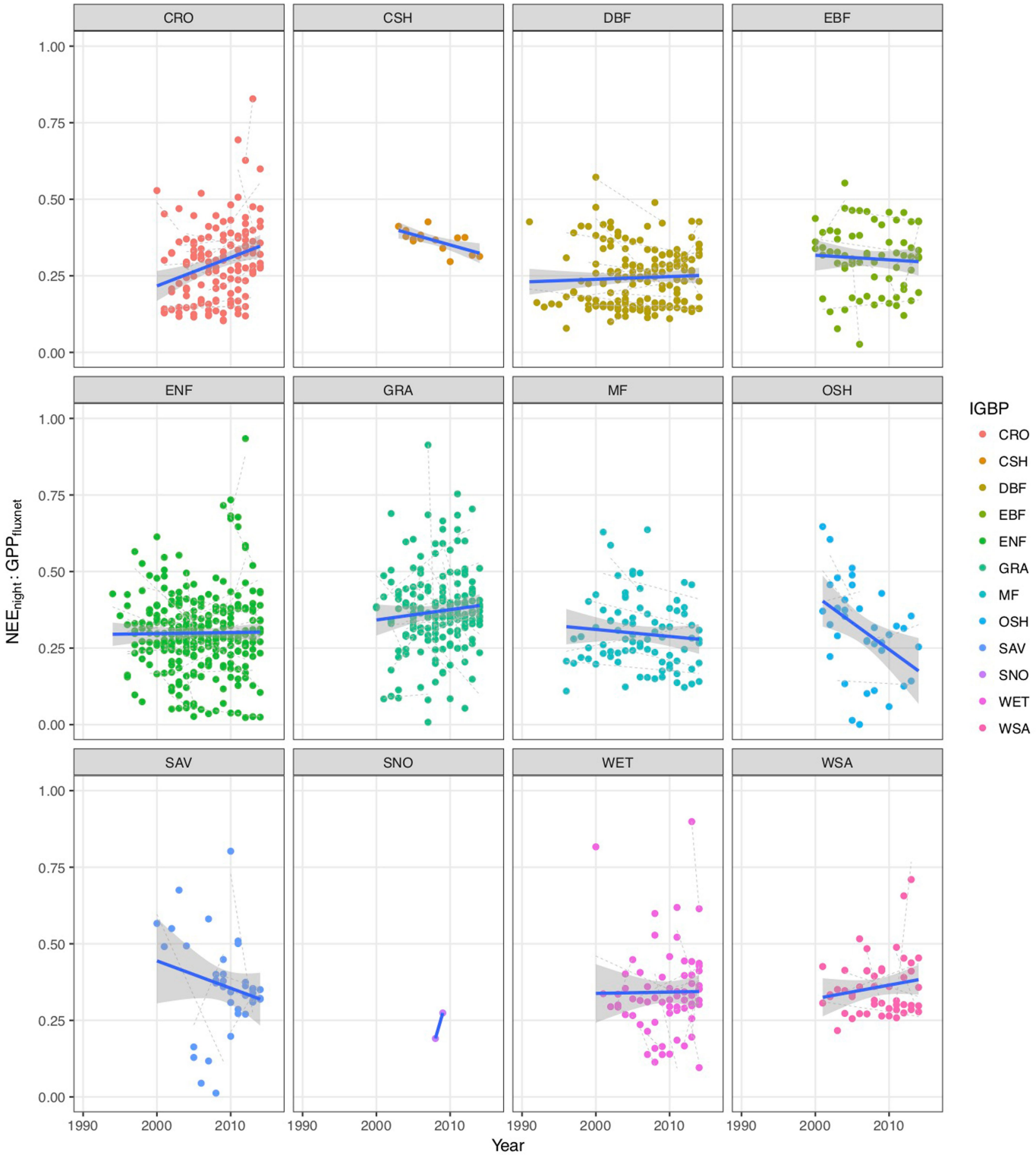
Extended Data Fig. 4 | Difference between FLUXNET GPP and remotely sensed GPP and SIF over time. Trends are not significant for the satellite data from GOME-2 SIF ($P = 0.22$), MODIS GPP ($P = 0.07$) or SCIMACHY SIF ($P = 0.49$). Blue lines (solid line is statistically significant, dashed line is non-significant) lines indicate least-squares trend. The temporal trend for MTE GPP decreases significantly ($P = 0.02$) over time, that is, MTE GPP increasingly underpredicts GPP measured by eddy covariance tower.

© 2018 Springer Nature Limited. All rights reserved.



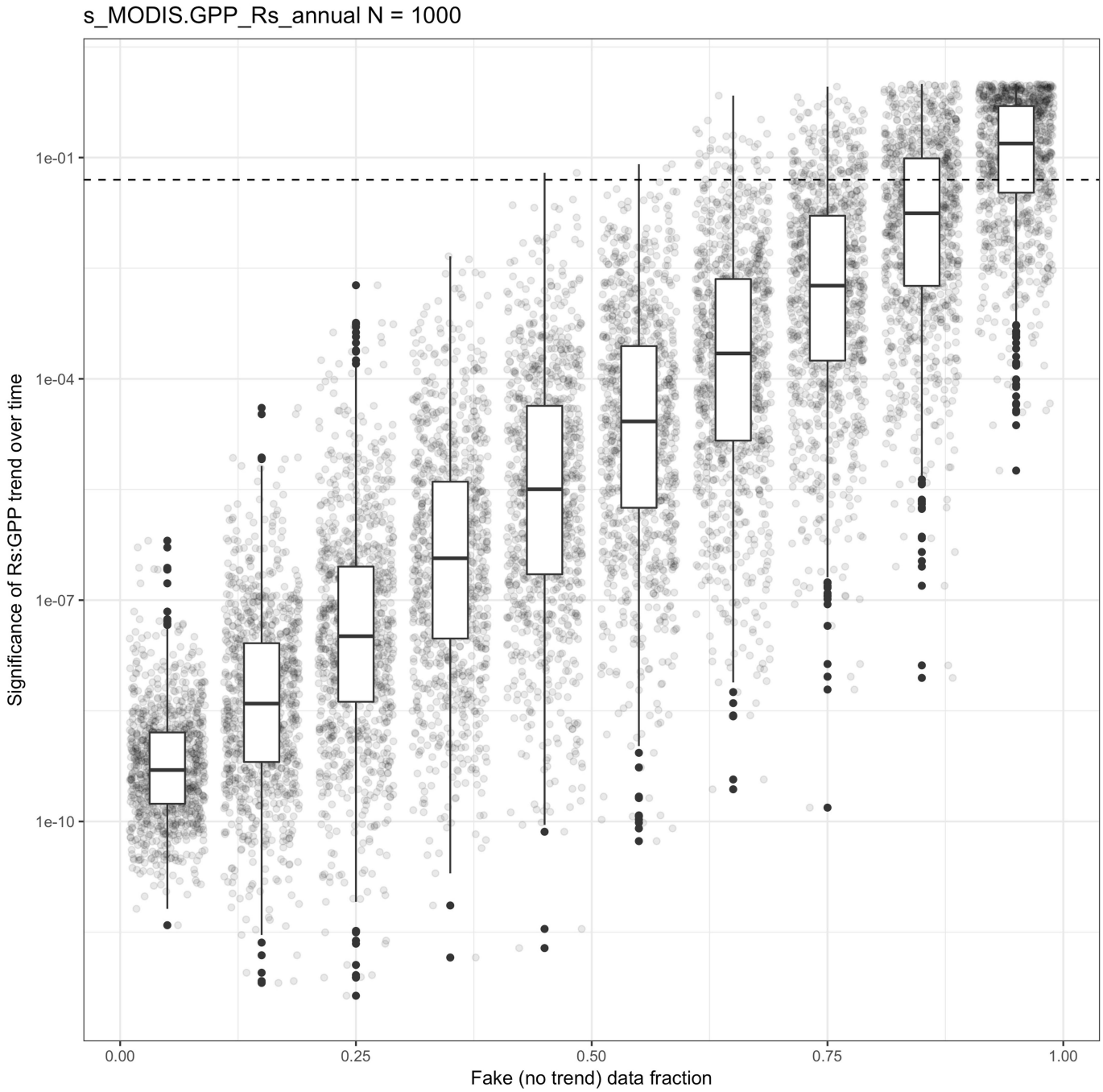
Extended Data Fig. 5 | Sensitivity analysis of how much GPP satellites would have to be missing to render the GPP–respiration trends shown in Fig. 2a non-significant. We ran a sensitivity analysis that assumed a conservative (that is, high) GPP trend of 0.5% yr^{-1}, and repeatedly re-fitted the Fig. 2a models assuming that satellites missed 0%, 10%, 20%,...100% of this gain. This figure plots the 'missed' percentage (*x* axis) versus the *P*-value of the linear regression (*y* axis) for each respiration flux (R_H and R_S) and GPP or SIF data product ('GPPadjusted', that is, adjusted for the assumed 'missed' percentage). Even if satellites are missing 100% of this large assumed GPP increase, most of the temporal trends shown in Fig. 2a trends remain highly significant.

© 2018 Springer Nature Limited. All rights reserved.



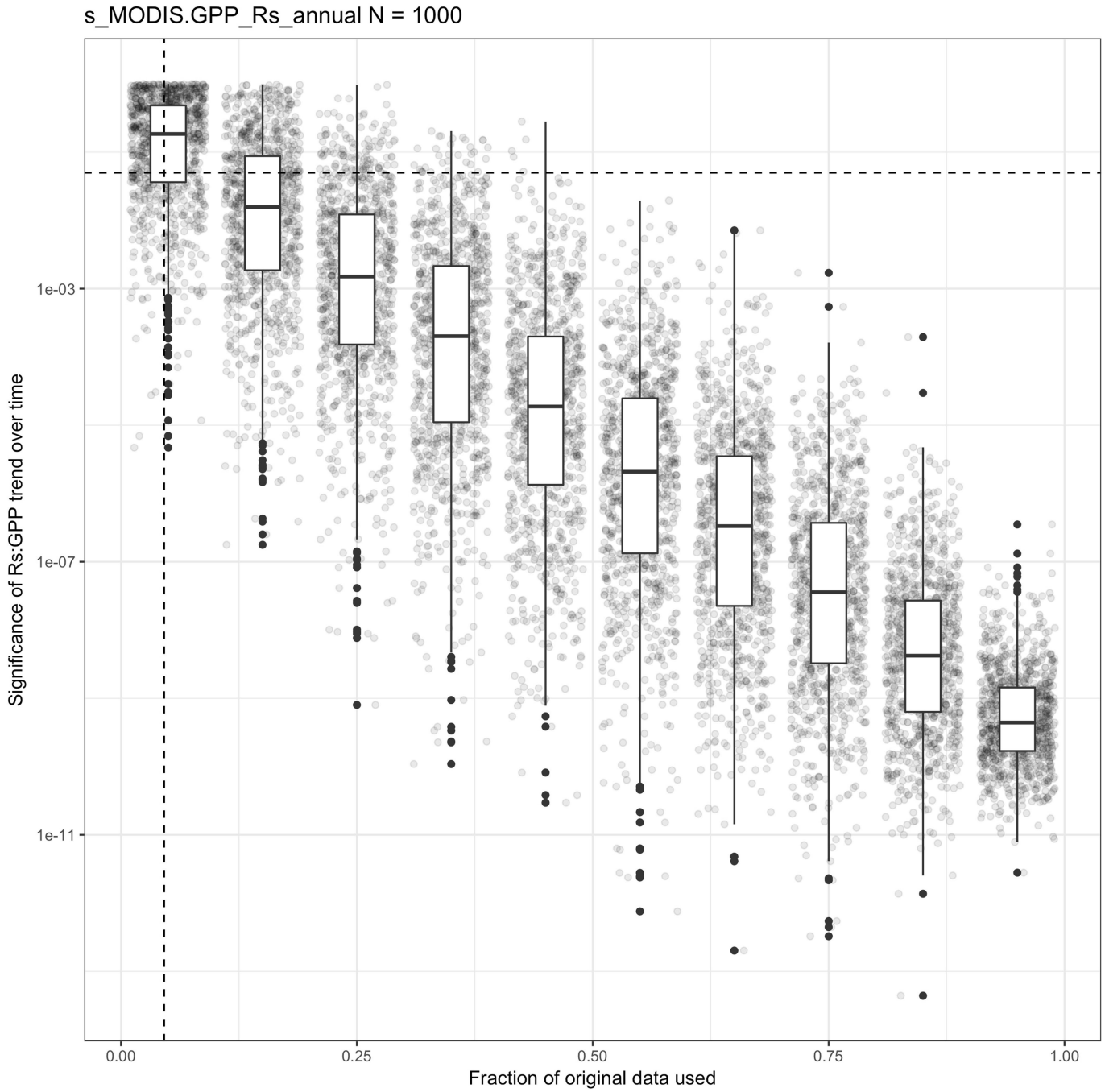
Extended Data Fig. 6 | Ratio of NEE_{night} to GPP in the FLUXNET 2015 'Tier 1' dataset. Panels show different International Geosphere-Biosphere Programme (IGBP) land cover type: crop (CRO), closed shrubland (CSH), deciduous broadleaf forest (DBF), evergreen broadleaf forest (EBF), evergreen needleleaf forest (ENF), grassland (GRA), mixed forests (MF), open shrublands (OSH), herbaceous savannahs (SAV), snow and ice (SNO), wetlands (WET) and woody savannahs (WSA).

© 2018 Springer Nature Limited. All rights reserved.



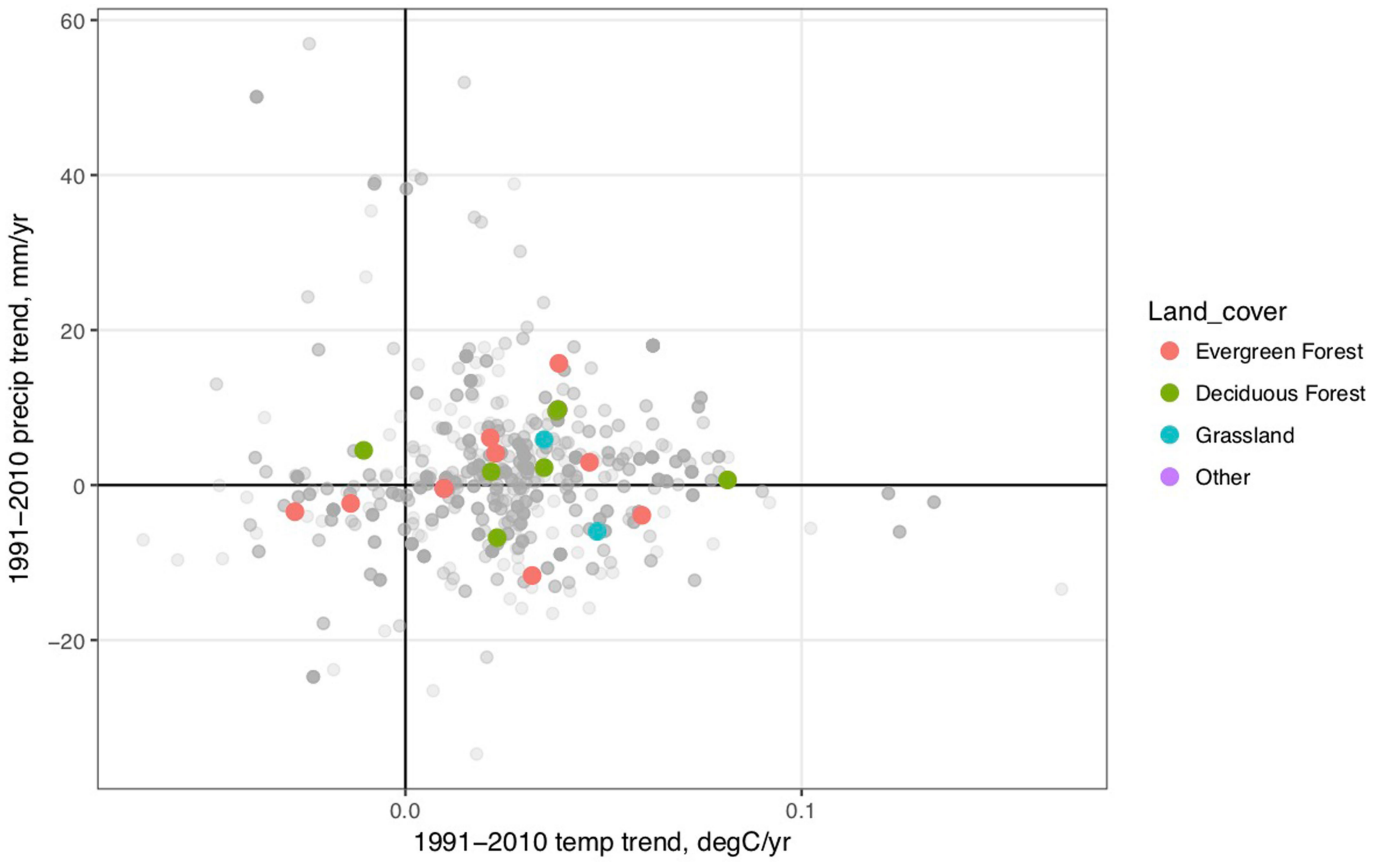
Extended Data Fig. 7 | Bootstrap analysis examining effect of ratio of significant to non-significant data on likelihood of observing a trend. What is the likelihood that the FLUXNET data subset (that included GPP measurements made in the same study site and year as site-specific RS measurements in the SRDB) is too small to detect signals of rising R_H, given site-to-site variability in climate and carbon dynamics? Each point is a different random draw from the original dataset, the comparison of R_S to MODIS GPP, with 1,000 bootstrap draws per fraction of artificial no-trend data. The horizontal dashed line is $P = 0.05$. Note log scale (to base e, which R defaults to) of y axis. Boxplots visualize five summary statistics (the median, the first and third quartiles, and two whiskers; see https://ggplot2.tidyverse.org/reference/geom_boxplot.html); points are semi-transparent to indicate data density at each point.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Bootstrap analysis examining effect of dataset size on likelihood of observing a trend. What is the likelihood that the $n = 106$ FLUXNET dataset is too small to detect signals of rising R_H, given site-to-site variability in climate and carbon dynamics? Each point is a different random draw from the original dataset, the comparison of R_S to MODIS GPP, with 1,000 bootstrap draws per fraction of no-trend data. The vertical dashed line shows the size of the original R_S:FLUXNET dataset, about 5% of the R_S:GPP_{MODIS} data, while the horizontal line is $P = 0.05$. Note log scale (to base e) of y axis. Boxplots visualize five summary statistics (the median, the first and third quartiles, and two whiskers; see https://ggplot2.tidyverse.org/reference/geom_boxplot.html); points are semi-transparent to indicate data density at each point.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Distribution of the longitudinal site data in climate change space relative to the main SRDB dataset. The longitudinal data (Fig. 2c, coloured dots below) in climate change space (1991–2010 HadCRUT4 changes in air temperature and precipitation) are shown over the distribution of the main dataset ($n = 1{,}852$, grey dots). The longitudinal data cover −0.03 to +0.08 °C yr^{-1} and −12 to +16 mm yr^{-1}; the main dataset covers a much broader climate change space of −0.07 to +0.17 °C yr^{-1} and −35 to +57 mm yr^{-1}.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Coefficients for general model[18] of relationship between R_S and R_H

Coefficient	Estimate	Std. Error	t	P
(Intercept)	-27.771	7.653	-3.629	<0.001
log(R_S)	0.867	0.033	26.044	<0.001
Year	0.014	0.004	3.653	<0.001

A model of form $\ln(R_H) \approx \ln(R_S)$*Year was fitted (following the protocol in ref. [18]) to the data shown in Fig. 1, with 'Year' being the numerical year of measurement. Coefficient name, value and standard error, *t* value and *P* value are shown. Model residual standard error = 0.335, adjusted $R^2 = 0.698$, $F_{2,317} = 396.6$, $P < 0.001$, AIC = 212.745.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0334-5

Blue boron-bearing diamonds from Earth's lower mantle

Evan M. Smith[1]*, Steven B. Shirey[2], Stephen H. Richardson[3], Fabrizio Nestola[4], Emma S. Bullock[5], Jianhua Wang[2] & Wuyi Wang[1]

Geological pathways for the recycling of Earth's surface materials into the mantle are both driven and obscured by plate tectonics[1–3]. Gauging the extent of this recycling is difficult because subducted crustal components are often released at relatively shallow depths, below arc volcanoes[4–7]. The conspicuous existence of blue boron-bearing diamonds (type IIb)[8,9] reveals that boron, an element abundant in the continental and oceanic crust, is present in certain diamond-forming fluids at mantle depths. However, both the provenance of the boron and the geological setting of diamond crystallization were unknown. Here we show that boron-bearing diamonds carry previously unrecognized mineral assemblages whose high-pressure precursors were stable in metamorphosed oceanic lithospheric slabs at depths reaching the lower mantle. We propose that some of the boron in seawater-serpentinized oceanic lithosphere is subducted into the deep mantle, where it is released with hydrous fluids that enable diamond growth[10]. Type IIb diamonds are thus among the deepest diamonds ever found and indicate a viable pathway for the deep-mantle recycling of crustal elements.

Type IIb diamonds—including the Hope, a renowned blue diamond—are mantle-derived minerals that contain boron at 0.01–10 p.p.m. levels and show a lack of nitrogen absorption in infrared spectroscopy[8]. Boron imparts their blue colour and p-type semiconductivity, although they may not always appear blue when boron concentrations are low or additional defects are present[9]. Because boron is a quintessential crustal element with a low concentration in Earth's mantle[11], blue diamonds and their formation have long been a geochemical enigma.

Type IIb diamonds have been recovered from worldwide localities, including southern and central Africa, India, South America and Borneo[9], having been brought to the surface in kimberlite volcanoes ranging in age from the 1.15-billion-year-old Premier pipe[12] to the ~90-million-year-old Letseng deposit[13]. They can reach large sizes, such as the 176.2-carat Brazilia diamond, the 122.5-carat rough diamond that yielded the (24.18-carat) Cullinan Dream, which was examined as part of this study, and the 112.5-carat rough diamond from which the Hope was cut[9].

The geological origin of blue diamonds has nevertheless remained unknown owing to their rarity (≤0.02% of mined diamonds; see Methods), high value and general lack of mineral inclusions. To overcome this problem, prospective samples were screened from the extensive grading operations of the Gemological Institute of America. Over two years, this approach allowed examination of 46 type IIb diamonds with inclusions, an invaluable suite for analysis (Extended Data Fig. 1).

Inclusions were characterized using Raman spectroscopy (Fig. 1) and were found to differ substantially from the common minerals found in diamonds from the cratonic lithosphere (<200 km), such as olivine and Cr-rich pyrope from peridotite or grossular-almandine-pyrope and omphacitic clinopyroxene from eclogite. Instead, the inclusion mineralogy is typical of super-deep diamonds originating from the mantle transition zone to the lower mantle (Extended Data Fig. 2)[14–17]. Inclusions in sublithospheric diamonds tend to destabilize during ascent in the mantle and break down to lower-pressure minerals, often unmixing into composite assemblages[14,17]. Many inclusions described here are multiphase assemblages, as is the case with previously studied inclusions in super-deep diamonds and their high-pressure experimental analogues[14,16,17]. It is implausible that the same multiphase assemblages could be coincidentally replicated by random sampling of lower-pressure mineral aggregates at shallower, lithospheric depths[14].

The most abundant inclusion identified, in 31 of 46 samples, was Ca-silicate dominated by $CaSiO_3$ walstromite, sometimes with larnite (β-Ca_2SiO_4) and other phases of $CaSiO_3$ composition (Extended Data Table 1). These inclusions are commonly interpreted as retrogressed $CaSiO_3$ perovskite (Ca-Pv)[15,17,18]. As retrogression of pure Ca-Pv alone should maintain a bulk Ca:Si ratio of 1, the presence of (Ca-rich) larnite in some inclusions may indicate that diamond growth occurred in a chemically evolving system with variable calcium enrichment, as seen in other super-deep diamonds[16,19,20].

Other observed inclusions also correspond to retrogressed high-pressure minerals (Extended Data Table 1). For example, inclusions of orthopyroxene, with sharp Raman spectra matching enstatite, and minor amounts of coexisting olivine are interpreted as retrogressed bridgmanite[14,15], the lower-mantle Mg-silicate perovskite phase. Multiphase inclusions containing ortho- or clinopyroxene, coexisting with jeffbenite ($Mg_3Al_2Si_3O_{12}$) or spinel ($(Mg,Fe)Al_2O_4$), are interpreted as aluminous bridgmanite[14], although some bearing clinopyroxene may represent retrogressed majoritic garnet[17].

One inclusion provides a convincing example of retrogressed majorite. Being fortuitously exposed on a facetted diamond, the two-phase assemblage of NaAl-clinopyroxene and jeffbenite (Fig. 1b) was confirmed with microanalysis by energy-dispersive X-ray spectroscopy (Extended Data Fig. 3) and was interpreted as a former low-Ca, high-Na majoritic garnet[17]. A separate inclusion of orthopyroxene in this diamond, interpreted as former bridgmanite, would then make a putative majorite–bridgmanite pair that would restrict its origin to within[17] ~660–750 km. Other observed inclusion phases are coesite (with accessory kyanite, interpreted as former stishovite) as well as ferropericlase (found as relatively small, brown inclusions; see Extended Data Fig. 4 and Supplementary Table 1).

Another diamond contains a multiphase inclusion dominated by nepheline and spinel, interpreted as former calcium-ferrite-type (CF) phase or possibly new aluminous (NAL) phase, which is compelling evidence of derivation from host rocks of basaltic composition at lower-mantle depths (Fig. 1)[14,16,17]. The same diamond also contains a multiphase inclusion of Fe carbide, Fe sulphide and wüstite (Extended Data Fig. 5, Supplementary Table 2) that does not correspond to a known mineral but may represent a former metallic melt similar to those recently discovered in (boron-lacking) CLIPPIR (Cullinan-like, Large, Inclusion-Poor, relatively Pure, Irregularly shaped and Resorbed) diamonds[21]. Three other type IIb samples also contain metallic-looking, magnetic inclusions that may be similar to those of CLIPPIR diamonds, although it should be stressed that these are a minor part of the type IIb

[1]Gemological Institute of America, New York, NY, USA. [2]Department of Terrestrial Magnetism, Carnegie Institution for Science, Washington, DC, USA. [3]Department of Geological Sciences, University of Cape Town, Rondebosch, South Africa. [4]Department of Geosciences, University of Padova, Padua, Italy. [5]Geophysical Laboratory, Carnegie Institution for Science, Washington, DC, USA. *e-mail: evan.smith@gia.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

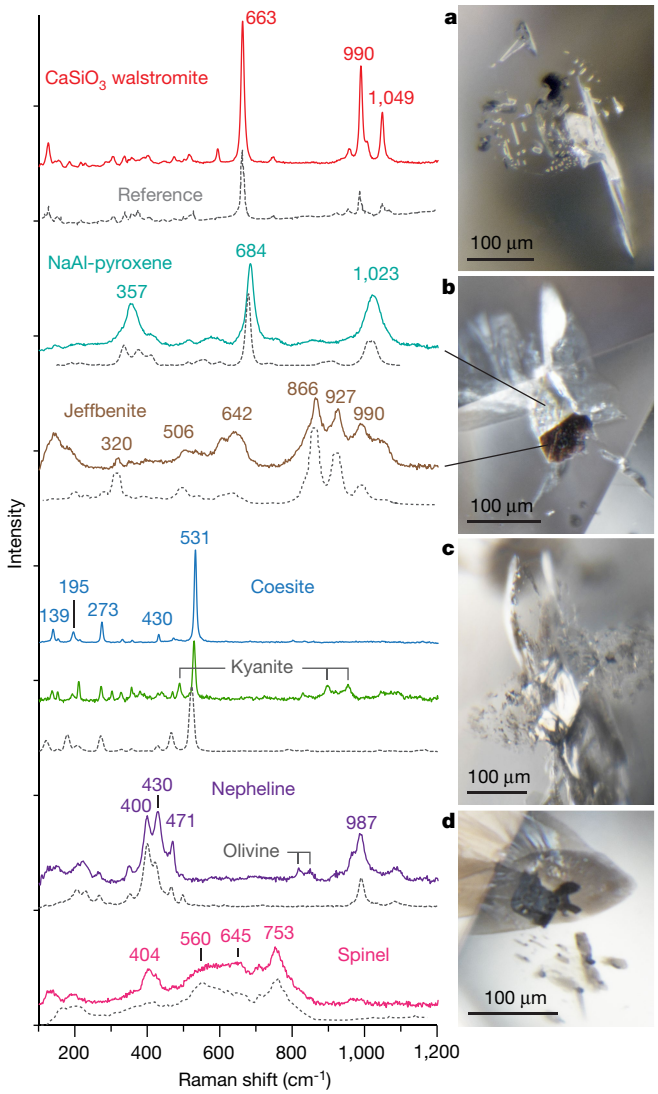


Fig. 1 | Selected Raman spectra of inclusions in type IIb diamonds. **a**, Former Ca-Pv, now $CaSiO_3$ walstromite, in sample 110205945970. **b**, Former majoritic garnet, now a composite of NaAl-pyroxene and jeffbenite (verified by scanning electron microscopy/energy-dispersive X-ray spectroscopy; see Extended Data Fig. 3) in sample 880000037816. **c**, Former stishovite, now coesite, in sample 101024478345. Also shown on the left is a composite coesite plus kyanite spectrum (green) from sample 890000180201. **d**, Former CF, now composite of nepheline and spinel (with CH_4 fluid, not shown), in sample 110208245246. Dashed lines are reference spectra of $CaSiO_3$ walstromite[19], jeffbenite[30], nepheline and spinel[14], plus omphacite R061129 and coesite X050094 from the RRUFF database. Spectra are stacked vertically for clarity. Source Data.

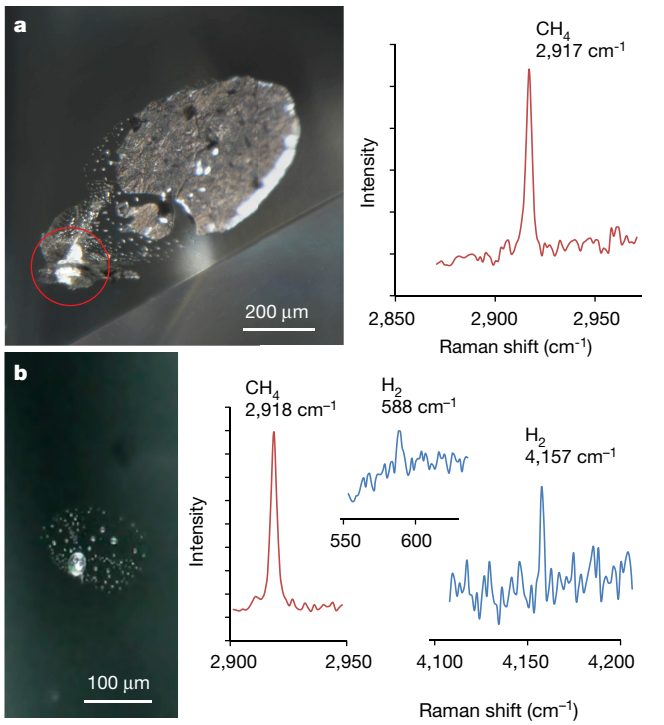


Fig. 2 | Inclusions jacketed by thin films of fluid CH_4 and H_2, revealed by Raman spectroscopy. **a**, $CaSiO_3$ walstromite (former Ca-Pv), with the main part of the inclusion circled, in sample 110208780369. **b**, Orthopyroxene (former bridgmanite) in sample 110208773706. Lobate sprays of small inclusions are thought to reflect expansion and proliferation of inclusion material into its own decompression crack.

diamond inclusion suite, whereas CLIPPIR diamonds are dominated by metallic Fe–Ni–C–S inclusions[21].

The observed inclusion mineralogy and the absence of typical silicate inclusions that characterize diamonds from the continental lithospheric mantle advocate for type IIb diamond growth in host rocks of basaltic-to-peridotitic bulk composition, consistent with the lithologies in subducted oceanic lithosphere reaching lower-mantle depths. Although some samples, particularly those containing Ca-Pv alone, may have grown in the mantle transition zone, the inclusion assemblages with former bridgmanite, ferropericlase and CF phase require an origin in the lower mantle[14,17].

In addition to inclusion mineralogy, a sublithospheric origin for blue diamonds is physically required by the extreme remnant pressure in some inclusions (see Methods). Using the pressure-induced shift in Raman spectral features, a 4.4 ± 0.1 GPa remnant pressure was measured in $CaSiO_3$ walstromite, which requires entrapment at a pressure of at least ~9 GPa in the mantle (we note that these entrapment pressures are severe underestimates due to diamond deformation; Extended Data Fig. 6)[22]. Similarly, a coesite inclusion with its main Raman peak shifted from 520.6 cm^{-1} up to 537.9 ± 0.5 cm^{-1} (Extended Data Fig. 6) indicates extreme remnant pressure, far exceeding the 4 GPa highest-pressure benchmark for coesites observed in lithospheric diamonds (see Methods). X-ray diffraction also reveals remnant pressures of ~1.8 GPa in ferropericlase inclusions, requiring a minimum entrapment pressure of 10.3 to 14.1 GPa, calculated at 1,200 K and 2,000 K, respectively, corresponding to a depth beyond 300 km, well below the base of the deepest continental lithospheric mantle keels[23].

Additional pressure observations come from the silicate inclusions in blue diamonds, which typically have large decompression cracks and often lobate sprays of tiny droplet-like satellite inclusions in healed fractures emanating from the main inclusion (Figs. 1, 2). These satellite inclusions formed during exhumation, as high internal inclusion stresses[20] relative to decreasing external confining pressure led to rupturing of the host diamond. Pervasive dislocation networks in type IIb diamonds (Extended Data Fig. 7) from plastic deformation and recovery[24] are also consistent with a high-temperature, sublithospheric mantle origin. Equivalent dislocation networks also appear in sublithospheric type IIa, CLIPPIR diamonds[21], but thus far have not been reported in diamonds from the cooler, shallower lithospheric mantle.

Methane, and often hydrogen, ($CH_4 \pm H_2$) were detected by Raman spectroscopy in 13 diamonds (28% of samples) as a thin fluid layer around one or more inclusions of varying mineralogy (Fig. 2). The fluid is a result of hydrogen escaping the inclusion and accumulating at the inclusion–host interface. When hydrogen is surrounded by diamond, it can form methane by reacting with the surrounding carbon. Similar $CH_4 \pm H_2$ fluid jackets have been found around metallic melt inclusions in natural and synthetic diamonds, formed when formerly dissolved atomic hydrogen diffused out of the inclusions upon cooling and decompression[21]. In the present mineral inclusions, these $CH_4 \pm H_2$ fluids are a strong indication that at least some of the retrograde

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

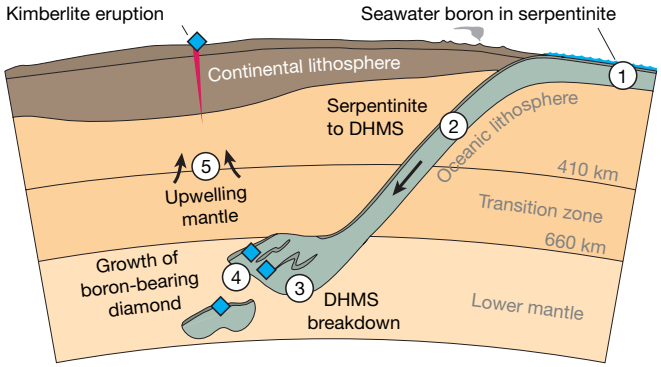


Fig. 3 | Formation of type IIb diamond. (1) Seafloor hydrothermal circulation causes serpentinization, introducing boron into oceanic lithosphere. (2) Subduction and metamorphism of serpentine to DHMS. (3) Breakdown of DHMS yields hydrous, boron-enriched fluid that migrates and evolves. It may gather carbon from the altered oceanic lithosphere. (4) Crystallization of boron-bearing diamond, triggered by redox reactions or in response to changing pressure, temperature or composition of the fluid. (5) Vertical transport may involve localized buoyancy associated with diamond-related metasomatism or an external mechanism such as a plume, with ultimate exhumation to the surface due to kimberlite volcanism.

mineral assemblage is hydrogen-saturated, implying that the original high-pressure minerals interacted with hydrous media.

Blue diamonds exhibit an inclusion mineralogy that requires diamond formation in host rocks of basaltic-to-peridotitic bulk composition in the lower mantle, much like inclusions described in other lower-mantle diamonds that have been linked to subducted oceanic lithosphere[14–18]. Inclusions of basaltic association, such as former stishovite or CF phase, clearly point to the involvement of subducted ocean crust[14,16], whereas those of peridotitic association, such as ferropericlase or low-Al bridgmanite, may be more closely related to the peridotitic portion of the oceanic lithospheric slab[15]. These subduction-related mineral assemblages are supported by a range of light-to-heavy carbon isotope signatures[16,25] in type IIb diamond, with three samples having a $\delta^{13}C$ of −13.4‰, −3.4‰ and −1.8‰ ($\delta^{13}C$ is the parts-per-thousand deviation of $^{13}C/^{12}C$ from the Pee Dee belemnite standard; see Supplementary Table 3), which complement three diamonds from a previous study with a $\delta^{13}C$ range of −20.8‰ to −14.5 ‰[26].

Boron is scarce in the convecting mantle, being 100 times more depleted compared to Earth's surface[5,11]. Its occurrence in blue diamonds implies an anomalously boron-enriched mantle source, especially considering that boron behaves incompatibly in diamond growth experiments. Although present observations do not rule out boron contribution from the ambient convecting mantle, the clear subducted-host-rock signatures strongly suggest that the boron is surface-derived, even if only small amounts of the full boron budget of the slab are subducted. Boron isotopic measurements may clarify its origin, but the prerequisite for non-destructive analysis, the lack of standards and the low boron concentrations in type IIb diamonds have been barriers to the use of current techniques. Nevertheless, support for a crustal boron source comes from the actual distribution of boron on Earth[11], the clear geophysical evidence of subducted slabs reaching the lower mantle[2] and the already established link between certain lower-mantle diamonds and subducted oceanic crust[14].

In deeply subducted oceanic lithosphere, the most likely boron source is metamorphosed serpentinized peridotite, with boron originally introduced by hydrothermal seawater circulation[4,5,7] (Fig. 3). Other boron carriers, such as sediments, organic matter or white micas in the ocean crust, are destabilized at relatively shallow depths, but serpentinized peridotite in the lithospheric-mantle portion of the slab can serve as an effective vehicle and is expected to be the largest boron reservoir within the slab[4,5,7]. Like water, boron is structurally hosted in serpentine[27]. Under the right conditions, cool subduction geotherms could successfully stabilize dense hydrous magnesium silicates (DHMSs) before serpentine breakdown, thereby creating a mineralogical pathway for boron, water and other components to reach the mantle transition zone and beyond[1,10,28]. The depth range of blue diamond genesis may therefore relate to DHMS breakdown, releasing hydrous diamond-forming fluid[10] as well as liberating boron (Fig. 3). Eventual transport of the blue diamonds to the surface is thought to involve upwelling mantle and kimberlite volcanism, as with other super-deep diamonds.

Blue diamonds point to a subduction-driven geochemical pathway extending from serpentinized oceanic lithosphere at Earth's surface to the lower mantle. This picture is consistent with key features of blue diamonds: mineral inclusions of basaltic-to-peridotitic bulk composition with affinity to subducted slabs[14]; high-pressure mineral assemblages indicating a depth of origin reaching the lower mantle, coinciding with the projected breakdown of DHMS minerals[10]; $CH_4 \pm H_2$ fluids suggestive of hydrous media; and, most remarkably, the boron content of the diamonds themselves. If metamorphosed serpentinite ultimately provides the boron for type IIb diamonds, as proposed here, then it is also implicated as a carrier of water to the deep mantle[28,29]. The recognition that blue diamonds originate from the lower mantle highlights a possible major pathway for ultra-deep water recycling on Earth.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0334-5.

Received: 12 January 2018; Accepted: 30 April 2018;
Published online 1 August 2018.

1. Kendrick, M. A. et al. Seawater cycled throughout Earth's mantle in partially serpentinized lithosphere. *Nat. Geosci.* **10**, 222–228 (2017).
2. Tackley, P. J. Mantle convection and plate tectonics: toward an integrated physical and chemical theory. *Science* **288**, 2002–2007 (2000).
3. Hofmann, A. W. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–229 (1997).
4. Leeman, W. P., Tonarini, S. & Turner, S. Boron isotope variations in Tonga-Kermadec-New Zealand arc lavas: implications for the origin of subduction components and mantle influences. *Geochem. Geophys. Geosyst.* **18**, 1126–1162 (2017).
5. Deschamps, F., Godard, M., Guillot, S. & Hattori, K. Geochemistry of subduction zone serpentinites: A review. *Lithos* **178**, 96–127 (2013).
6. Plank, T. & Langmuir, C. H. The chemical composition of subducting sediment and its consequences for the crust and mantle. *Chem. Geol.* **145**, 325–394 (1998).
7. Konrad-Schmolke, M., Halama, R. & Manea, V. C. Slab mantle dehydrates beneath Kamchatka—yet recycles water into the deep mantle. *Geochem. Geophys. Geosyst.* **17**, 2987–3007 (2016).
8. Gaillou, E., Post, J. E., Rost, D. & Butler, J. E. Boron in natural type IIb blue diamonds: chemical and spectroscopic measurements. *Am. Mineral.* **97**, 1–18 (2012).
9. King, J. M. et al. Characterizing natural-color type IIb blue diamonds. *Gems Gemol.* **34**, 246–268 (1998).
10. Harte, B. Diamond formation in the deep mantle: the record of mineral inclusions and their distribution in relation to mantle dehydration zones. *Mineral. Mag.* **74**, 189–215 (2010).
11. Grew, E. S. Boron: from cosmic scarcity to 300 minerals. *Elements* **13**, 225–229 (2017).
12. Wu, F. Y. et al. In situ U-Pb age determination and Sr-Nd isotopic analysis of perovskite from the Premier (Cullinan) kimberlite, South Africa. *Chem. Geol.* **353**, 83–95 (2013).
13. Skinner, E. & Truswell, J. in *The Geology of South Africa* (eds Johnson, M. R. et al.) 651–659 (Geological Society of South Africa, Johannesburg, 2006).
14. Walter, M. J. et al. Deep mantle cycling of oceanic crust: evidence from diamonds and their mineral inclusions. *Science* **334**, 54–57 (2011).
15. Stachel, T., Harris, J. W., Brey, G. P. & Joswig, W. Kankan diamonds (Guinea) II: lower mantle inclusion parageneses. *Contrib. Mineral. Petrol.* **140**, 16–27 (2000).
16. Thomson, A. et al. Origin of sub-lithospheric diamonds from the Juina-5 kimberlite (Brazil): constraints from carbon isotopes and inclusion compositions. *Contrib. Mineral. Petrol.* **168**, 1081 (2014).
17. Harte, B. & Hudson, N. C. F. In *Proc. 10th International Kimberlite Conference Vol.* **1** (eds Pearson, D. G. et al.) 235–253 (Springer, New Delhi, 2013).
18. Nestola, F. et al. $CaSiO_3$ perovskite in diamond indicates the recycling of oceanic crust into the lower mantle. *Nature* **555**, 237–241 (2018).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

19. Brenker, F. E. et al. Detection of a Ca-rich lithology in the Earth's deep (> 300 km) convecting mantle. *Earth Planet. Sci. Lett.* **236**, 579–587 (2005).
20. Anzolini, C. et al. Depth of formation of $CaSiO_3$-walstromite included in super-deep diamonds. *Lithos* **265**, 138–147 (2016).
21. Smith, E. M. et al. Large gem diamonds from metallic liquid in Earth's deep mantle. *Science* **354**, 1403–1405 (2016).
22. Anzolini, C. et al. Depth of formation of super-deep diamonds: Raman barometry of $CaSiO_3$-walstromite inclusions. *Am. Mineral.* **103**, 69–74 (2018).
23. Angel, R. J., Mazzucchelli, M. L., Alvaro, M. & Nestola, F. EosFit-Pinc: a simple GUI for host-inclusion elastic thermobarometry. *Am. Mineral.* **102**, 1957–1960 (2017).
24. Hanley, P. L., Kiflawi, I. & Lang, A. R. On topographically identifiable sources of cathodoluminescence in natural diamonds. *Phil. Trans. R. Soc. Lond. A* **284**, 329–368 (1977).
25. Stachel, T., Harris, J., Aulbach, S. & Deines, P. Kankan diamonds (Guinea) III: $d^{13}C$ and nitrogen characteristics of deep diamonds. *Contrib. Mineral. Petrol.* **142**, 465–475 (2002).
26. Milledge, H. J. et al. Carbon isotopic variation in spectral type II diamonds. *Nature* **303**, 791–792 (1983).
27. Pabst, S. et al. Evidence for boron incorporation into the serpentine crystal structure. *Am. Mineral.* **96**, 1112–1119 (2011).
28. Ohtani, E., Litasov, K., Hosoya, T., Kubo, T. & Kondo, T. Water transport into the deep mantle and formation of a hydrous transition zone. *Phys. Earth Planet. Inter.* **143–144**, 255–269 (2004).
29. Pearson, D. G. et al. Hydrous mantle transition zone indicated by ringwoodite included within diamond. *Nature* **507**, 221–224 (2014).
30. Nestola, F. et al. Tetragonal Almandine-Pyrope Phase, TAPP: finally a name for it, the new mineral jeffbenite. *Mineral. Mag.* **80**, 1219–1232 (2016).

Acknowledgements This research was supported by a GIA Liddicoat Postdoctoral Research Fellowship to E.M.S. Support to S.B.S. and F.N. was provided by the Deep Carbon Observatory (DCO). The European Research Council supported F.N. (INDIMEDEA, number 307322). Sincere thanks to K. S. Moe, T. Moses, M. Breeding, U. D'Haenens-Johansson, P. Johnson, K. Smit, J. Liao, S. Persaud, E. Myagkaya, A. Balter and B. Torres for analytical/logistical assistance, to N. Renfro for the micrograph of Fig. 1a, to Ascot Diamonds for lending rough samples and to M. Alvaro for discussions about geobarometry.

Reviewer information *Nature* thanks E. Gaillou and T. Stachel for their contribution to the peer review of this work.

Author contributions E.M.S. led the research, characterized the samples, conducted Raman analyses, interpreted results and wrote the initial manuscript. S.B.S. and S.H.R. contributed scientific interpretations and substantive manuscript writing. F.N. conducted X-ray diffraction and assisted with geobarometry. E.S.B. provided support for electron microprobe analysis. J.W. conducted mass spectroscopy for carbon. W.W. helped guide the project and ensured access to samples and analytical resources.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0334-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0334-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to E.M.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Raman spectroscopy. Raman spectroscopy was performed at the Gemological Institute of America (GIA), New York, with a Renishaw InVia Raman Microscope, using a grating with 1,800 lines mm^{-1}, a 50× lens with a numerical aperture of 0.55, and the 514.5 nm (and 488 nm for additional analyses) output of an argon-ion laser set to 150 mW output power. Calibration was done using the 520.5 cm^{-1} Raman peak of a polished silicon wafer.

Energy dispersive X-ray spectroscopy, electron microscopy and cathodoluminescence imaging. Qualitative scanning electron microscopy/energy-dispersive X-ray spectroscopy (SEM-EDS) examination was carried out at GIA, New York, using a Zeiss EVO MA10 SEM system with an Oxford X-Max^N 20 EDS detector, in variable-pressure mode with chamber pressure of 20 Pa and electron beam conditions of 10–20 kV and 1 nA. The same instrument was used to capture secondary-electron and backscatter images of exposed inclusions, with a chamber pressure of 20 Pa and electron beam conditions of 15 kV and 1 nA, as well as cathodoluminescence images (black and white, panchromatic), with a chamber pressure of 20 Pa and electron beam conditions of 15 kV and 0.5 nA.

Electron microprobe analysis. Electron microprobe analyses were performed at the Geophysical Laboratory of the Carnegie Institution for Science on a JEOL JXA-8530F field-emission electron probe. The operating conditions were 15 kV and 30 nA. Samples were coated with carbon. For the carbide analyses, Fe_7C_3, NiS and an Al-bearing enstatite glass were used as standards. The Fe_7C_3 standard was a synthetic high-pressure iron carbide sample prepared at the University of Michigan[31].

X-ray diffraction. X-ray diffraction analyses were done in the Department of Geosciences, University of Padova using a custom instrumentation setup including a high-brilliance MoKα micro-beam source operating at 50 kV and 0.8 mA, a Rigaku Oxford Diffraction SuperNova goniometer and a Dectris Pilatus 200K detector. The setup allows measurement of inclusions of very limited size, down to 5–10 μm, that are still in situ within diamonds. Although two samples were examined (110208245246 and 110208425476), owing to the difficulty in aligning the microbeam with individual inclusions, only the latter sample successfully produced reflections from inclusion phases.

Mass spectroscopy. Carbon isotopic analysis was done at the Department of Terrestrial Magnetism of the Carnegie Institution for Science on a CAMECA 6F ion microprobe using an extreme energy filtering method. A 10 kV primary Cs beam of 5 nA was focused to ~15 μm and the beam centre was rastered across a 20 μm × 20 μm area on the sample, held at −4,650 V. Secondary ions of ^{12}C and ^{13}C were extracted at −5 kV with an energy window of 100 V. The ^{12}C ions were counted for 1 s and the ^{13}C ions for 5 s, and counting was repeated 200 times. The 2σ counting precision for ^{13}C was about 0.5 ‰. The diamond standards N198, 4139, 1013, 'Mao' and JWH-NS were used to determine the instrumental mass fractionation and drift before and after sample analyses.

Prevalence of type IIb diamonds. Type IIb diamonds are recognized as being relatively rare, but their rarity is seldom—if ever—defined quantitatively. To determine a quantitative figure, a review was conducted for a sample set of 13.8 million natural gem-quality diamonds submitted to GIA. This sample set excludes synthetic diamonds, which are often of type IIb and might otherwise skew the results. Within the sample set of natural gem diamonds, 0.02% were recorded as being type IIb. We note that this figure is considered a robust cross-section of natural gem diamonds, but it comes with some uncertainty. A small number of the diamonds surveyed may be duplicates, submitted for grading multiple times within the sample period; this may disproportionately affect the count of type IIb samples, especially for higher-valued blue gems. Also, the tally of type IIb diamonds may be inflated because they are more likely to be submitted to GIA for grading because of their colour, and thus their higher gem value. Conversely, a small proportion of type IIb diamonds may not have been officially recorded as type IIb in the database used, and by default would be counted as non-type IIb. Lastly, and most importantly, the above value is expressly for gem diamonds, and does not account for industrial goods. Given the colour, morphology and clarity characteristics of known type IIb diamonds, it is considered unlikely that many industrial-quality natural diamonds are of type IIb. Because about 70%–80% of mined diamonds are of industrial quality, the proportion of all diamonds (gem plus industrial) that are of type IIb may be as low as 0.004%.

Samples. A list of all samples examined and the inclusion phases observed within them is given in Extended Data Table 1. We note that a diamond may contain multiple inclusions of the same mineralogy and that the list does not necessarily reflect the number of inclusions found in each sample. However, the abundance of a certain kind of inclusion within individual diamonds compares very well with the frequency of those inclusions across the whole suite. More important is the inclusion assemblage and the potential host rock paragenesis that it portrays. A visual overview of the sample suite is given in Extended Data Fig. 1.

Metallic Fe–S–C–O multiphase inclusion. Sample 110208245246 contains an opaque inclusion (inclusion B) whose mineralogy was unclear from the initial Raman analysis. Subsequent polishing to expose this inclusion to examine it by electron microscopy revealed it to be a Fe–S–C–O multiphase inclusion, with three main phases: Fe sulphide, Fe carbide, and Fe oxide (Extended Data Fig. 5). These phases were also analysed using an electron microprobe (Supplementary Tables 1 and 2). The Fe sulphide matches pyrrhotite by composition, $Fe_{1-x}S$, with minor variable surface tarnish smeared across the polished surface of the Fe sulphide, thought to be an artefact of the polishing process (see mottled texture of Fe sulphide phase in the oxygen X-ray map; Extended Data Fig. 5). The Fe oxide composition matches well with wüstite, $Fe_{1-x}O$. The Fe carbide phase was verified to contain carbon, but quantification was complicated by the additional carbon signal from the carbon coat (and potentially the underlying and neighbouring host diamond material). The carbide phase is tentatively interpreted as an Eckstrom–Adcock carbide (ideally Fe_7C_3), although cohenite (Fe_3C) is not ruled out without further analysis. This metallic Fe–S–C–O multiphase inclusion is similar to, but still distinct from, the Fe–Ni–C–S inclusions recently reported from CLIPPIR diamonds (a variety of large, inclusion-poor, often type IIa, sublithospheric diamonds)[21]. In the samples examined, metallic Fe–S–C–O inclusions are a relatively rare occurrence, as opposed to Fe–Ni–C–S inclusions, which are the most common inclusion type seen in CLIPPIR diamonds.

Sample containing ferropericlase (plus olivine and nyerereite), verified by X-ray diffraction. Sample 110208425476 contains multiple dark inclusions. These inclusions are unusual compared to those in the rest of the samples because they are very irregular in shape, distribution and abundance within the sample. These dark inclusions are accompanied by smaller (down to micrometre size) inclusions defining curviplanar features interpreted as healed cracks. Where these features intersect the polished surfaces, the polish is interrupted slightly, as would be expected for the slight crystallographic misalignment of a healed crack. In cathodoluminescence imaging, these healed cracks appear as bright lines, which are intimately connected within the diamond's strong overall dislocation network pattern (Extended Data Fig. 4). The cathodoluminescence image reveals complex features texturally consistent with intense bulk plastic/brittle deformation.

Initial Raman spectroscopy revealed some inclusions to be (or at least contain) ferropericlase, according to two broad Raman bands at around 650–690 cm^{-1} and 190 cm^{-1}, which correspond to ferropericlase in other diamonds in our experience. Later, polishing down to some of the inclusions revealed examples of ferropericlase, which were analysed with an electron microprobe (Supplementary Table 1).

Many inclusions, however, did not yield any clear Raman signal. The ferropericlase inclusions (Extended Data Fig. 4) may have been modified during diamond deformation, especially if large cracks developed that allowed melt or fluid to penetrate and alter or augment the inclusion mineralogy. Examination with X-ray diffraction (University of Padova) revealed the presence of not only ferropericlase (unit cell edge length, 4.209(8) Å; volume, 74.5(3) $Å^3$; uncertainties are estimated standard deviations per convention), but also olivine (unit cell dimensions, 4.771(4) Å, 10.266(6) Å, 6.004(3) Å; angles, 90.0°, 90.0°, 90.0°; volume, 294.0(3) $Å^3$), and nyerereite, $Na_2Ca(CO_3)_2$ (observed d-spacings are 3.22, 6.42, 1.79, 1.52 and potentially others that overlap with diamond reflections). Some additional reflections were observed that could not be conclusively assigned: 1.44 (very intense), 2.04 (very intense, not overlapped with diamond), 2.40 (medium intensity), 2.86 (low intensity) and very-low-intensity peaks at 1.84, 1.70 and 1.75. It is possible that the first four reflections (1.44, 2.04, 2.40 and 2.86) correspond to ringwoodite, a polymorph of olivine stable in the lower half of the mantle transition zone. The relatively large measured olivine unit cell certainly suggests Mg-poor compositions (Mg/(Mg+Fe) < 0.80). Accounting for even a modest remnant pressure of 0.5–0.6 GPa[32,33] would imply an olivine composition of Fo_{73} (Mg/(Mg+Fe) = 0.73), which is far from typical olivine-inclusion compositions, but overlaps with the composition reported for a ringwoodite inclusion with $Mg/(Mg+Fe) = 0.75^{0.24}_{0.21}$ (ref. 29). It is therefore considered possible that this sample may contain both olivine from inverted ringwoodite, as well as preserved ringwoodite. Again, it should be noted that the healed cracks raise the possibility that some inclusion material could have been introduced post-growth, meaning that ferropericlase may not necessarily have been in equilibrium with other phases, nyerereite and olivine. For this reason, and given that ferropericlase was observed in other samples, only ferropericlase is discussed and considered as a primary mineral in the main text.

Raman- and X-ray-diffraction-based inclusion barometry. Within diamonds, the pressure inside an inclusion is often elevated. Although the extreme confining pressure of the deep mantle is removed as the diamond is carried to the surface, the inclusion remains confined within the diamond host and may still have some remnant pressure. The expected remnant pressure is primarily dependent on the inclusion mineral species and depth (pressure and temperature conditions) of entrapment. In reality, most inclusions have at least some reduction in their remnant pressure due to diamond deformation (brittle/plastic) relieving some built-up pressure. Determining the remnant pressure in an inclusion can give a physical indication of the minimum depth of entrapment, constraining the depth of diamond growth[20,22,23,32,34–40].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Raman spectroscopy is based on the energy associated with inducing molecular vibrations, which in some instances can be sensitive to stresses acting upon the material. This means that the Raman spectrum of some materials exhibits measurable, predictable changes under pressure. The Raman peaks of minerals, for example, can be slightly shifted from their expected values when the minerals are under pressure.

In the suite of type IIb diamonds, two inclusions were found to preserve extreme remnant pressure. The first inclusion is a $CaSiO_3$ walstromite inclusion, which although it is interpreted to be an inversion product from Ca-Pv, still preserves remnant pressure that can help retrace the final stages of exhumation from the mantle. The inclusion, in sample 110203744064, has its 977.3 cm^{-1} peak (unconfined position) shifted upwards to $1{,}000.4 \pm 0.5$ cm^{-1} (Extended Data Fig. 6), which corresponds to a remnant pressure of 4.4 ± 0.1 GPa, using the mineral- and peak-specific experimental calibration factor[22] of 5.16 ± 0.9 cm^{-1} GPa^{-1}. Assuming any reasonable mantle temperature within 1,200–2,000 K, this remnant pressure would have required a minimum inclusion entrapment pressure of ~9 GPa (approximately 240–280 km deep)[22].

Similarly, in sample 890000180198, the main Raman peak of coesite is shifted from its unconfined value of 520.6 cm^{-1} to 537.9 ± 0.5 cm^{-1} (Extended Data Fig. 6). Using the pressure calibration factor[41] of 2.9 cm^{-1} GPa^{-1} gives a nominal remnant pressure of 6.0 ± 0.4 GPa, but this is not considered a reliable pressure determination. Other peaks in the spectrum do not exhibit the same magnitude of pressure-induced shift, which suggests anisotropic stresses. Caution is needed in the interpretation of remnant pressure in coesite owing to the complex effects of anisotropy, anomalies in its elastic behaviour, plus the lack of reliable high-pressure and low-to-high-temperature experiments for coesite. A non-cubic mineral trapped in a cubic (almost elastically isotropic) host such as diamond will be subject to anisotropic strains during exhumation from the mantle and will develop deviatoric stresses. Currently there is no reliable method among existing analytical techniques and hydrostatic calibrations to determine strains on minerals subject to deviatoric stresses and interpret them in terms of pressure. X-ray diffraction could be a useful method to supplement Raman spectroscopy for this purpose, but unfortunately it cannot be performed on small inclusions to the required level of precision and accuracy. These factors limit the interpretation of the Raman results from this inclusion. A minimum entrapment pressure is therefore not derived here. Nevertheless, the remnant pressure suggested by the main peak of coesite far exceeds those of other coesite inclusions found in lithospheric diamonds, which reach upwards of 3.6–4.3 GPa[42–44].

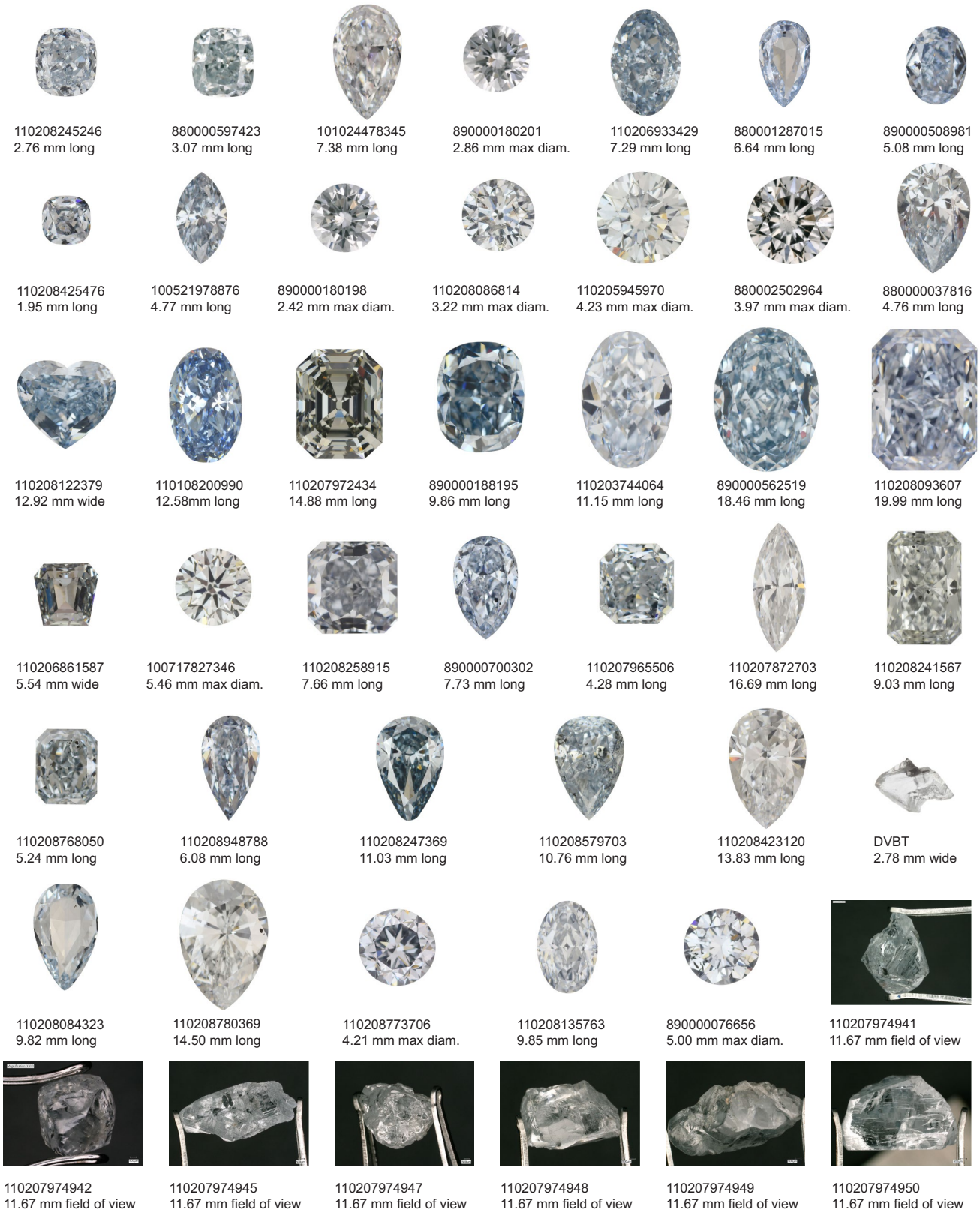
In addition to Raman spectroscopy, X-ray diffraction revealed high remnant pressure inside ferropericlase inclusions in one of the studied type IIb diamonds, sample 110208425476. X-ray diffraction permitted the calculation of the unit cell (74.5(3) $Å^3$), which, when combined with the composition obtained later by an electron microprobe ($Mg/(Mg+Fe) = 0.91$), gives a remnant pressure of 1.8 GPa. At mantle temperatures of 1,200–2,000 K, this is estimated to require entrapment at pressures of at least 10.3–14.1 GPa[23], not accounting for the stress relieved by brittle and plastic strains.

For all three inclusions, $CaSiO_3$ walstromite, coesite and ferropericlase, the estimated entrapment conditions are considered as minimum pressures. Much of the inclusion pressure has been relieved by brittle and plastic deformation of the surrounding diamond. Taken together, these results require that the diamonds originated from below the lithosphere.

Data availability. All relevant data are presented in Extended Data Figs. 1–7, Extended Data Table 1 and Supplementary Information Tables 1–3. Original spectral data and electron microprobe data are available from the corresponding author.

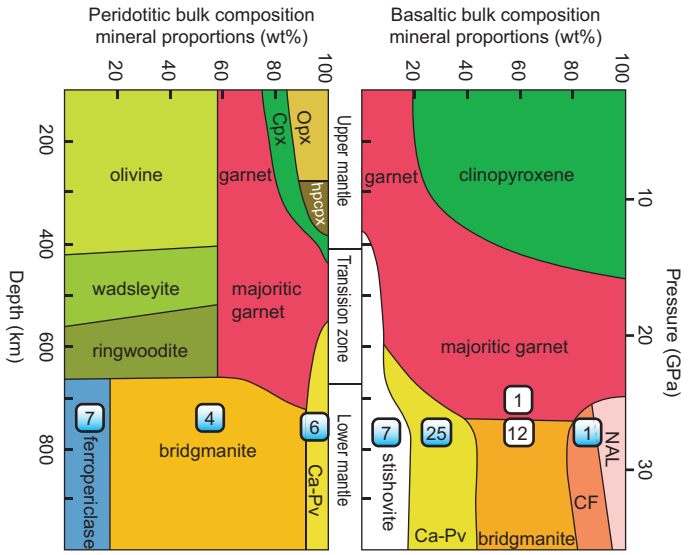
31. Liu, J., Li, J. & Ikuta, D. Elastic softening in Fe_7C_3 with implications for Earth's deep carbon reservoirs. *J. Geophys. Res. Solid Earth* **121**, 1514–1524 (2016).
32. Howell, D., Wood, I. G., Nestola, F., Nimis, P. & Nasdala, L. Inclusions under remnant pressure in diamond: a multi-technique approach. *Eur. J. Mineral.* **24**, 563–573 (2012).
33. Nestola, F. et al. First crystal-structure determination of olivine in diamond: composition and implications for provenance in the Earth's mantle. *Earth Planet. Sci. Lett.* **305**, 249–255 (2011).
34. Mazzucchelli, M. L. et al. Elastic geothermobarometry: corrections for the geometry of the host-inclusion system. *Geology* **46**, 231–234 (2018).
35. Angel, R. J., Alvaro, M., Miletich, R. & Nestola, F. A simple and generalised P–T–V EoS for continuous phase transitions, implemented in EosFit and applied to quartz. *Contrib. Mineral. Petrol.* **172**, 29 (2017).
36. Gonzalez-Platas, J., Alvaro, M., Nestola, F. & Angel, R. EosFit7-GUI: a new graphical user interface for equation of state calculations, analyses and teaching. *J. Appl. Cryst.* **49**, 1377–1382 (2016).
37. Angel, R., Alvaro, M., Nestola, F. & Mazzucchelli, M. Diamond thermoelastic properties and implications for determining the pressure of formation of diamond-inclusion systems. *Russ. Geol. Geophys.* **56**, 211–220 (2015).
38. Angel, R., Nimis, P., Mazzucchelli, M., Alvaro, M. & Nestola, F. How large are departures from lithostatic pressure? Constraints from host–inclusion elasticity. *J. Metamorph. Geol.* **33**, 801–813 (2015).
39. Angel, R. J., Alvaro, M. & Gonzalez-Platas, J. EosFit7c and a Fortran module (library) for equation of state calculations. *Z. Kristallogr. Cryst. Mater.* **229**, 405–419 (2014).
40. Angel, R. J., Mazzucchelli, M. L., Alvaro, M., Nimis, P. & Nestola, F. Geobarometry from host-inclusion systems: the role of elastic relaxation. *Am. Mineral.* **99**, 2146–2149 (2014).
41. Hemley, R. J. in *High-Pressure Research in Mineral Physics Vol.* **39** (eds Manghnani, M. H. & Syono, Y.) 347–359 (American Geophysical Union, Washington DC, 1987).
42. Sobolev, N. V. et al. Fossilized high pressure from the Earth's deep interior: the coesite-in-diamond barometer. *Proc. Natl Acad. Sci. USA* **97**, 11875–11879 (2000).
43. Smith, E. M., Kopylova, M. G., Frezzotti, M. L. & Afanasiev, V. P. Fluid inclusions in Ebelyakh diamonds: evidence of CO_2 liberation in eclogite and the effect of H_2O on diamond habit. *Lithos* **216-217**, 106–117 (2015).
44. Smith, E. M., Vendrell, C. & Johnson, P. Coesite inclusions with filaments in diamond. *Gems Gemol.* **52**, 410–412 (2016).
45. D'Ippolito, V., Andreozzi, G. B., Bersani, D. & Lottici, P. P. Raman fingerprint of chromate, aluminate and ferrite spinels. *J. Raman Spectrosc.* **46**, 1255–1264 (2015).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



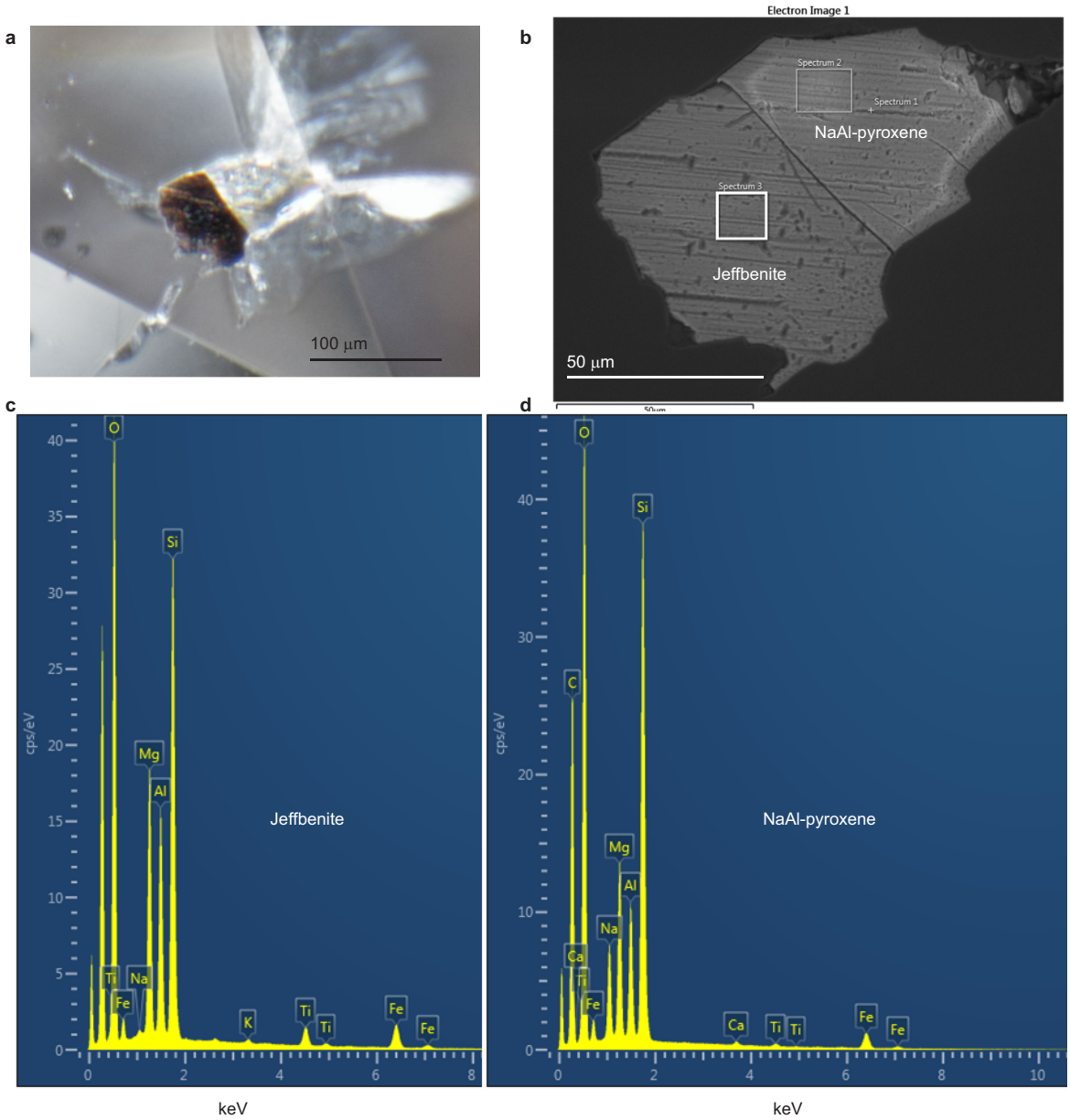
Extended Data Fig. 1 | Suite of 46 type IIb diamonds studied. Images are not to scale. Refer to noted dimensions (max diam., maximum diameter).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



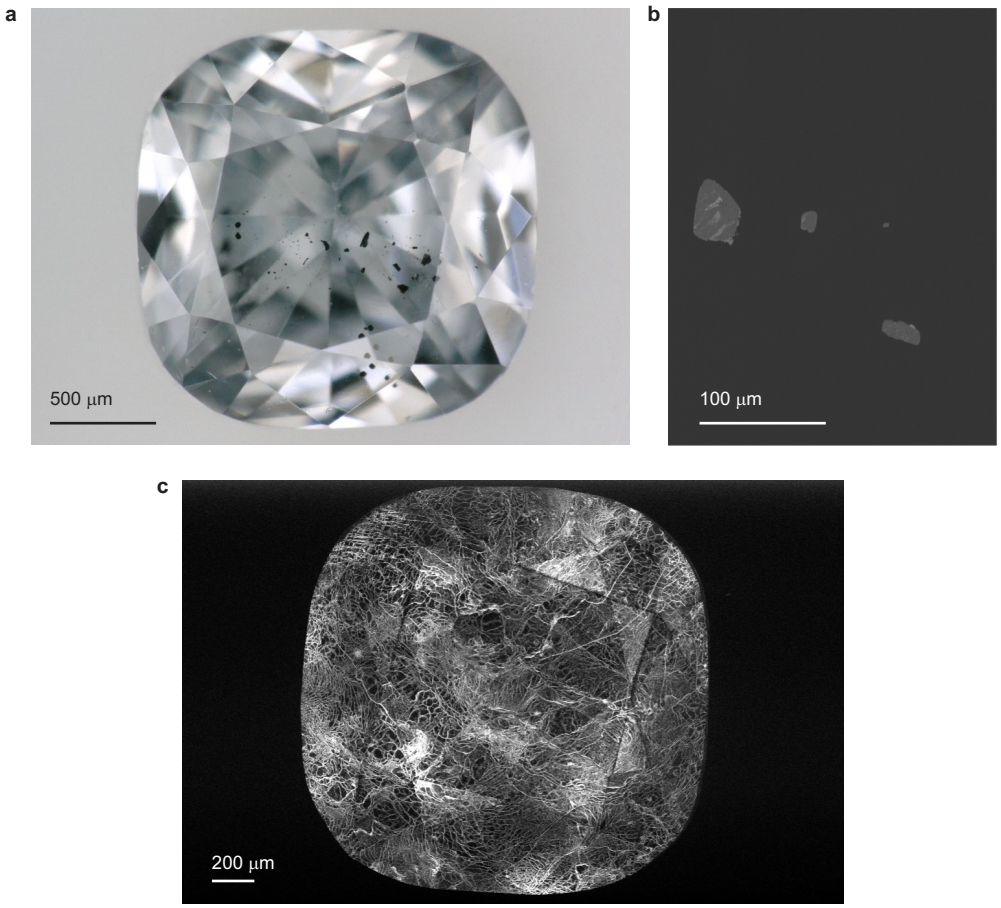
Extended Data Fig. 2 | Mineralogy of mantle rocks with peridotitic and basaltic bulk composition as a function of depth. Numbers in boxes denote the number of diamonds observed with inclusions of the given phase, and blue shading in the boxes indicates that a thin fluid $CH_4 \pm H_2$ jacket was found with the phase. We note that the division of samples between the left and right panel is for illustrative purposes, and in reality some samples (for example, with Ca-Pv alone) are not firmly categorized. See Extended Data Table 1 for a breakdown of inclusions by sample. Adapted from ref. 17.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



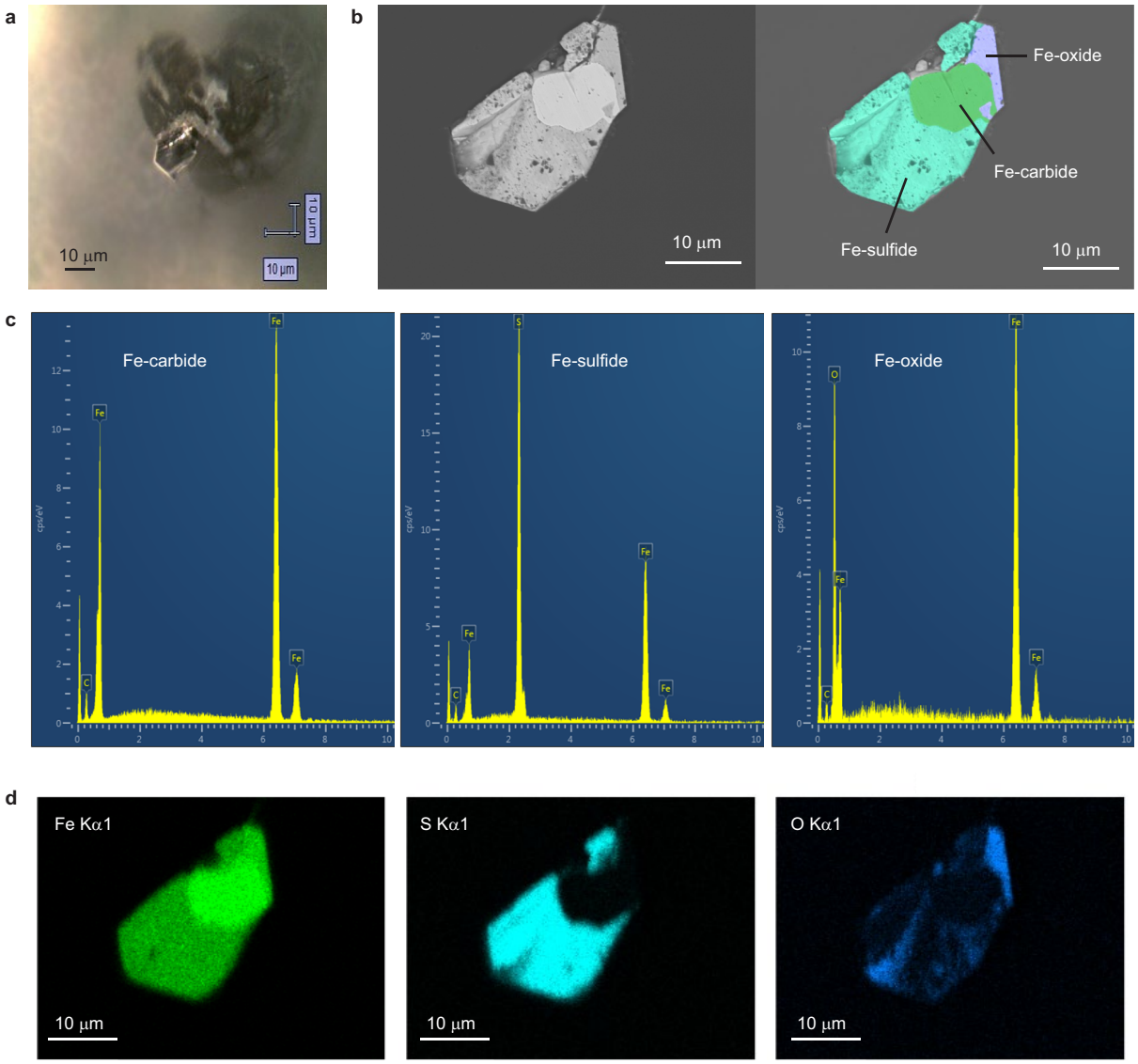
Extended Data Fig. 3 | Multiphase inclusion interpreted as former low-Ca, high-Na majoritic garnet. a, Optical microscope image of the inclusion, exposed on a polished facet of sample 880000037816. **b**, Secondary-electron image of the same inclusion, grooved with nearly horizontal polishing lines. **c**, **d**, EDS spectra of the two phases, consistent with their Raman identification as jeffbenite and NaAl-pyroxene (monoclinic, with composition between enstatite and jadeite). High Na content suggests a metabasaltic paragenesis, while low Ca content may reflect Ca partitioning into coexisting Ca-Pv at the base of the mantle transition zone or the uppermost lower mantle.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



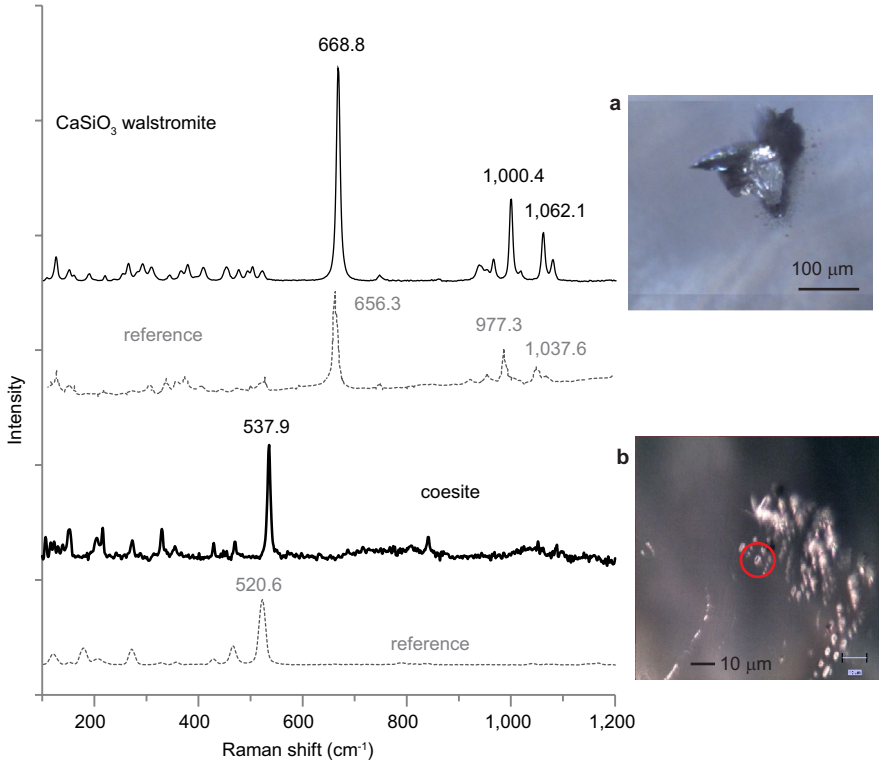
Extended Data Fig. 4 | Diamond sample 110208425476. a, Optical microscope image of the whole diamond, showing multiple dark inclusions of ferropericlase. **b**, Polishing down the table facet slightly exposed this group of four ferropericlase inclusions, shown here in an electron backscatter image. The smeared texture on the largest inclusion is a small amount of iron inadvertently deposited on the surface during polishing on a conventional cast iron scaife. **c**, Cathodoluminescence image of the whole diamond, revealing a complex dislocation network pattern, with interspersed healed fractures, that records a combination of both plastic and brittle deformation.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



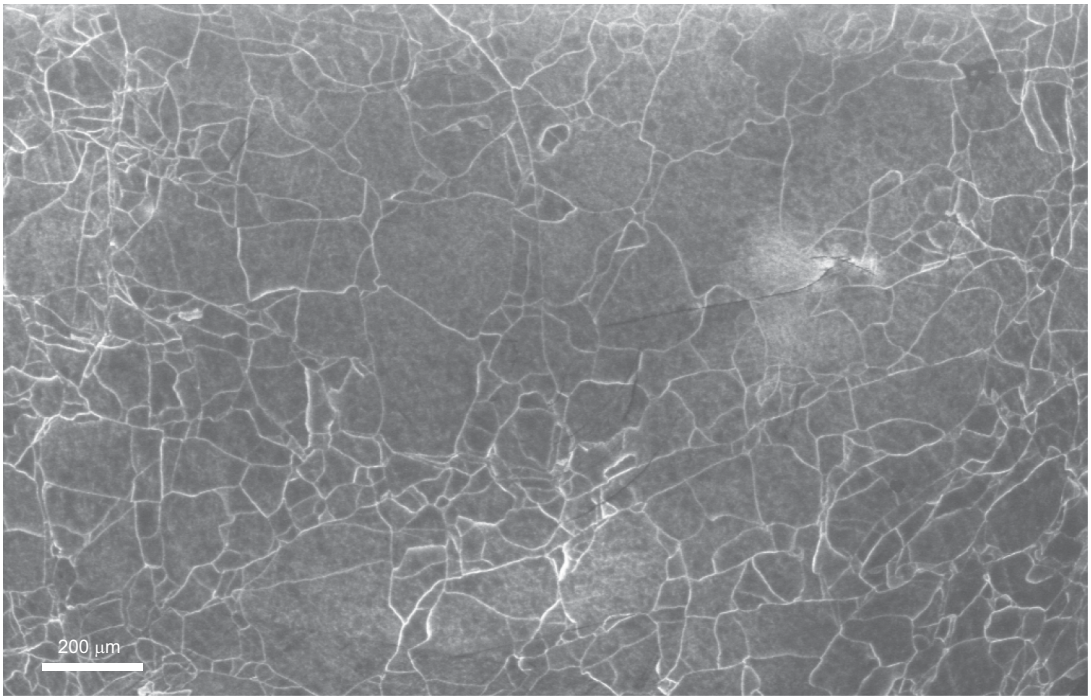
Extended Data Fig. 5 | Multiphase Fe–S–C–O metallic inclusion in sample 110208245246 (inclusion B). **a**, Optical microscope image of inclusion B, while still contained within the diamond host. **b**, Electron backscatter image of the inclusion, after polishing to expose a cross-section through it. The three main phases are colour-coded in the right panel. **c**, X-ray spectra obtained with EDS, showing the qualitative elemental composition of each of the three phases. Carbon is present in all spectra owing to the diamond host, diamond particles embedded in the polished surface (black specks, especially in the sulphide phase), as well as the carbon inherent to the Fe-carbide phase. **d**, X-ray elemental maps obtained with EDS, showing the spatial variation in signal in the region of Fe, S and O peaks (Kα1). Sulphur delineates the Fe-sulphide phase. Oxygen marks the Fe-oxide phase, while also showing the variable oxidation/tarnish layer on the sulphide portion of the inclusion.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Two inclusions exhibiting a large pressure-induced shift in Raman features. a, $CaSiO_3$ walstromite (thought to be former Ca-Pv) in sample 110203744064, inclusion A, with the three main peaks shifted to higher wavenumbers compared to a zero-pressure reference spectrum. This inclusion also contains CH_4. **b**, Coesite (SiO_2, thought to be former stishovite) in sample 890000180198. The inclusion analysed (circled) is about 2 μm wide and lies in a planar lobate healed crack, along with other related 'satellite' inclusions that presumably surrounded a nucleus coesite inclusion that was polished away when this diamond was facetted. Neighbouring coesite inclusions in **b** also have high, but variable, remnant pressures, as reflected by the Raman spectra. Reference spectra are from ref. 19 and RRUFF-X050094, and zero-pressure reference peak positions are taken from refs[22,41].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Dislocation network pattern in sample 110208245246, as seen in panchromatic cathodoluminescence. Each of the bright web-like lines are made up of many dislocations, and these cathodoluminescent boundaries surround darker, low-strain domains. The dark curved feature on the right of the centre is a crack (not healed).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Suite of 46 type IIb diamond samples and inclusion summary

Sample	Mass (carats)	Observed single- or multi-phase inclusions, interpreted as: [Ca-Pv, Ca-silicate perovskite] *[Bridgmanite, Mg-silicate perovskite] [Stishovite] [CF, calcium ferrite-type phase] [Ferropericlase] [sodic majoritic garnet] †[Bridgmanite, aluminum bearing] [metallic alloy/mixture] (Mineral abbrev. in footnote)
Basaltic compositional association, containing CF, Stishovite, or sodic majoritic garnet		
110208245246	0.08	[Wal], [Opx + Jeffbenite + Spinel + Ilmenite], [Nepheline + Olivine + Spinel + **CH_4**], [FeS + Fe-carbide + FeO]
880000597423	0.13	[Pyroxene + Spinel + Corundum + Olivine], [Coesite + trace Px + **CH_4**]
101024478345	0.57	[Cpx + Opx + Jeffbenite], [Coesite], ‡[possible sulfide], §[small unidentified opaque], §[Calcite, in healed cracks]
890000180201	0.08	[Pyroxene + Jeffbenite], [Coesite + Kyanite]
110206933429	1.01	[Wal + Larnite], [Cpx + Jeffbenite + Spinel], [Coesite], ‡[possible sulfide], ‡[Perovskite, may be $CaSiO_3$ or $CaTiO_3$]
100521978876	0.13	[Wal + Larnite + Wollastonite], [Coesite]
890000180198	0.05	[Wal + Larnite], [Coesite]
880002502964	0.24	[Cpx + Jeffbenite], [Coesite], [small unidentified opaque]
880000037816	0.21	[Wal], [NaAl-pyroxene + Jeffbenite; examined with SEM-EDS], [Opx], ‡[possible sulfide]
May associate with basaltic or peridotitic bulk compositions, not firmly categorized		
110205945970	0.27	[Wal + Larnite], [Opx + Jeffbenite], ‡[possible sulfide]
110208086814	0.13	[Wal + Wollastonite], [Pyroxene + Jeffbenite + Spinel + Ilmenite + Olivine], [Opx + Spinel], ‡[possible sulfide]
100717827346	0.61	[Cpx + Opx + Spinel + Olivine] Raman clearly shows two spectrally/spatially distinct pyroxenes
110208579703	2.01	[Wal + Larnite + **CH_4**], [Opx], [Unidentified opaque, magnetic + **CH_4**], §[lobe of small unidentified inclusions + **CH_4**, magnetic, suspected metallic alloy]
890000188195	3.46	[Wal + Larnite + **CH_4**]
890000700302	0.61	[Wal + Larnite + **CH_4 + H_2**], §[lobe of small unidentified inclusions + **CH_4**, magnetic, suspected metallic alloy]
110203744064	2.70	[Wal + **CH_4**]
110208093607	24.18	[Wal + Larnite] *Cullinan Dream diamond, 122.52 carat rough mass*
890000562519	17.09	[Wal + Larnite]
110108200990	3.81	[Wal + Larnite]
110208258915	2.26	[Wal + Larnite]
110207965506	0.46	[Wal + Larnite]
110207872703	2.46	[Wal + Larnite]
110208241567	2.15	[Wal + Larnite]
110206861587	0.73	[Wal + Larnite]
110208948788	0.32	[Wal + Larnite]
110208247369	2.08	[Wal]
110207972434	10.67	[Wal]
DVBT	0.03	[Wal], §[small Fe-Ni-S inclusions, examined with SEM-EDS, in healed lobate crack]
110208423120	4.06	[Wal + Pseudowollastonite]
880001287015	0.35	[Wal + Larnite + Pseudowollastonite], [Unidentified opaque, magnetic + **CH_4 + H_2**]
890000508981	0.30	[Opx + Jeffbenite]
110208122379	6.08	[Unidentified opaque, 655 band like wüstite]
110208768050	0.42	‡[possible sulfide]
110208084323	1.02	‡[possible sulfide]
110208135763	1.71	[graphitic fracture rosette hides inclusion; suspected Wal or sulfide based on appearance]
110207974950	2.97	[graphitic fracture rosette hides inclusion; suspected Wal or sulfide based on appearance]
Peridotitic compositional association, with fPer and/or Opx having distinctly sharp enstatite Raman spectrum and no Al-phases		
110208773706	0.32	[Wal + Larnite], [Opx + **CH_4 + H_2**]
110207974945	0.78	[Wal], [Opx + Olivine + **CH_4 + H_2**, plus weak 253+376 may be lepidocrocite γ-FeOOH], [fPer], [fPer + unidentified opaque + **CH_4**], §[lobe of small unidentified inclusions + **CH_4**]
110207974941	0.92	[Opx + **CH_4**], [Unidentified opaque, magnetic + **CH_4 + H_2**, suspected metallic alloy]
890000076656	0.48	[Wal + Larnite], [Opx + Olivine + **CH_4**]
110207974949	2.80	[Wal + Larnite], [fPer]
110207974948	2.17	[Wal + Larnite], [fPer], §[lobe of small unidentified inclusions + **CH_4**, plus weak 253+376 may be lepidocrocite γ-FeOOH]
110208780369	5.02	[Wal + Pseudowollastonite + **CH_4**], [fPer], [fPer + minor sulfate or phosphate, and sulfide phases]
110207974947	1.45	[fPer + Unidentified opaque], [graphitic fracture, inclusion nucleus not visible]
110207974942	2.31	[fPer + Unidentified opaque], §[Dolomite, in healed cracks]
110208425476	0.03	§[fPer, irregularly shaped inclusions, pervaded by healed fractures; XRD shows fPer + Olivine + nyerereite]

Wal, $CaSiO_3$ walstromite; Cpx, clinopyroxene; Opx, orthopyroxene (resolvable doublet in the 660–680 cm^{-1} region); fPer, ferropericlase. Pyroxene not specified as clino- or ortho-pyroxene was not confidently subcategorized from its asymmetric main peak at 660–690 cm^{-1}. The identified spinel has Raman spectral features like those of $MgAl_2O_4$ and $FeAl_2O_4$[45], a prominent peak at ~750 cm^{-1}, and other peaks matching spinel reported in other retrogressed CF-phase and Al-bridgmanite inclusions[14]. The possible sulphide phase has a Raman spectrum resembling arsenopyrite and other As or Sb sulphides, with a Raman spectrum having weak, sharp, variable peaks in the region 100–350 cm^{-1}. Pseudowollastonite is a high-temperature $CaSiO_3$ polymorph, found here as an accessory within three calcium silicate inclusions (main peaks: 985, 580, 374 and 142 cm^{-1}).

*Bridgmanite identification based on the presence of orthopyroxene, with sharp Raman peaks matching enstatite ($Mg_2Si_2O_6$) with a clearly resolved doublet in the 660–680 cm^{-1} region. The absence of jeffbenite/spinel suggests low-Al bridgmanite.

†Aluminous bridgmanite is the likely precursor for these multi-phase assemblages. However, some of these inclusions may actually represent original majoritic garnet if the pyroxene phase were to be identified as NaAl-pyroxene, which can be difficult to resolve in some cases. It should be stressed that even with chemical analysis, interpretation of these multi-phase retrogressed inclusions is not straightforward, and the significance of jeffbenite in retrogressed assemblages remains a matter of debate[17].

‡Small (<10 μm, often <5 μm) and rare inclusion, regarded as a minor accessory inclusion.

§Inclusions in healed cracks, of mantle origin but trapped or modified post-growth.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0352-3

Hurricane-induced selection on the morphology of an island lizard

Colin M. Donihue[1,2]*, Anthony Herrel[2,3,4], Anne-Claire Fabre[2], Ambika Kamath[5], Anthony J. Geneva[1], Thomas W. Schoener[6], Jason J. Kolbe[7] & Jonathan B. Losos[8]

Hurricanes are catastrophically destructive. Beyond their toll on human life and livelihoods, hurricanes have tremendous and often long-lasting effects on ecological systems[1,2]. Despite many examples of mass mortality events following hurricanes[3–5], hurricane-induced natural selection has not previously been demonstrated. Immediately after we finished a survey of *Anolis scriptus*—a common, small-bodied lizard found throughout the Turks and Caicos archipelago—our study populations were battered by Hurricanes Irma and Maria. Shortly thereafter, we revisited the populations to determine whether morphological traits related to clinging capacity had shifted in the intervening six weeks and found that populations of surviving lizards differed in body size, relative limb length and toepad size from those present before the storm. Our serendipitous study, which to our knowledge is the first to use an immediately before and after comparison[6] to investigate selection caused by hurricanes, demonstrates that hurricanes can induce phenotypic change in a population and strongly implicates natural selection as the cause. In the decades ahead, as extreme climate events are predicted to become more intense and prevalent[7,8], our understanding of evolutionary dynamics needs to incorporate the effects of these potentially severe selective episodes[9–11].

Hurricanes bring death and destruction to the ecosystems in their path. Their effects are myriad—decimating populations[3,4], reshuffling plant and animal communities[1,2,5,12], and fundamentally altering ecosystem cycles[2,13]. Long-term demography studies have provided numerous examples of substantial mortality events in plants[14], sponges[15], land snails[4], stick insects[4], lizards[16], birds[5] and monkeys[3] due to hurricanes. It remains an open question, however, whether hurricanes cause selective mortality on the basis of phenotypic traits or are instead so indiscriminately destructive that survival is random with regard to phenotype.

The fall of 2017 was an extraordinary season for Atlantic storms with three record-breaking events: Hurricanes Harvey, Irma and Maria. Four days before Hurricane Irma made landfall in the Turks and Caicos Islands, we finished surveying the morphology of *A. scriptus* on two small islands (Pine Cay and Water Cay) in the archipelago (Fig. 1). Hurricane Irma hit the islands with sustained 265-kph winds and, two weeks later, Hurricane Maria followed with sustained 200-kph winds.

Six weeks after our initial survey and three weeks after Hurricane Maria, we returned to Pine Cay and Water Cay, resurveyed the same transects on both islands, caught a sample of the surviving lizards and measured their morphology (Methods). Decades of previous research on *Anolis* species have demonstrated that the size of their toepads and the length of their limbs are related to habitat use and locomotor performance. We therefore tested whether the mean toepad surface area[17] and limb lengths[18] of the surviving lizards were larger than those of the lizards in the populations that we initially surveyed, as these increases are predicted to improve clinging ability[17,18].

We found parallel shifts in limb and toepad morphology between the pre- and post-hurricane populations on both islands (see Supplementary Information for complete results and model output). A multivariate analysis of covariance on all limb components revealed that the morphology of the post-hurricane lizard populations differed significantly from the pre-hurricane lizard populations ($F_{146} = 18.278$, $P < 0.0001$), and that the post-hurricane shifts were parallel on both islands—there was no hurricane × island interaction; $F_{146} = 1.377$, $P = 0.1833$. The multivariate analysis of covariance structure coefficients indicate relative femur length most significantly distinguished the pre- and post-hurricane populations (see Supplementary Information).

Post hoc analyses examining traits individually indicate that the surviving populations of both islands had proportionately longer humeri (relative humerus length: $\beta \pm$ s.e.: 0.03 ± 0.008; $t_{159} = 3.64$; $P = 0.0004$; Fig. 2)—the average humerus length increased by 1.8%, despite a significant 1.4% decrease in mean body size (snout-to-vent length (SVL): $\beta \pm$ s.e.: 1.20 ± 0.484; $t_{159} = -2.483$; $P = 0.0141$). By contrast, we observed a significant 6% decrease in relative femur length across both populations after the hurricane ($\beta \pm$ s.e.: -0.05 ± 0.009; $t_{159} = -5.92$; $P < 0.0001$; Fig. 2) and a decrease of 4.6% in the length of the longest toe (relative longest toe length: $\beta \pm$ s.e.: -0.03 ± 0.009; $t_{159} = -3.75$; $P = 0.0002$; Supplementary Information).

On average, the post-hurricane lizards of both islands had larger toepads on both their forelimbs and hindlimbs (relative forelimb toepad area: $\beta \pm$ s.e.: 0.13 ± 0.017; $t_{158} = 7.93$; $P < 0.0001$; relative hindlimb toepad area: $\beta \pm$ s.e.: 0.10 ± 0.015; $t_{158} = 6.819$; $P < 0.0001$; Fig. 2), which correspond to increases in population means of 9.2% and 6.1% for forelimb and hindlimb toepad area, respectively. We also found a parallel sex × hurricane interaction in body size ($\beta \pm$ s.e.: 2.87 ± 0.949; $t_{159} = 3.027$; $P = 0.003$)—on average, male SVL decreased by 4.3% whereas female SVL increased by 0.9%.

Despite the overall trend for parallel shifts, which was evident in both the multivariate analysis of covariance and many of the morphological traits, we found that forelimb toe length showed a differing response between the two islands (hurricane × island interaction: $\beta \pm$ s.e.: -0.05 ± 0.025; $t_{159} = -2.11$; $P = 0.036$), decreasing by 3.9% on Pine Cay and increasing by 2.2% on Water Cay. In addition to the parallel and non-parallel changes detected for many traits, we did not detect a difference between pre- and post-hurricane populations in the length of any other segments of the limbs, nor in the number of lamellar scales on the forelimb or hindlimb toepads (Supplementary Information).

We next considered what might be responsible for the parallel shifts in phenotypes in the two populations. Two lines of evidence suggest that natural selection favoured individuals able to survive the hurricanes. First, if the hurricanes caused directional selection, then the survivors should exhibit reduced trait variation after the hurricanes[6]. We tested this prediction by calculating the variance in each of the measurements that showed a significant shift in their mean after the

[1]Harvard University, Department of Organismic and Evolutionary Biology, Cambridge, MA, USA. [2]UMR 7179 CNRS/MNHN, Département Adaptations du Vivant, Paris, France. [3]Ghent University, Department of Biology, Evolutionary Morphology of Vertebrates, Ghent, Belgium. [4]University of Antwerp, Department of Biology, Functional Morphology, Antwerp, Belgium. [5]University of California, Department of Ecology, Evolution, and Marine Biology, Santa Barbara, CA, USA. [6]University of California, Department of Evolution and Ecology, Davis, CA, USA. [7]University of Rhode Island, Department of Biological Sciences, Kingston, RI, USA. [8]Washington University, Department of Biology, St. Louis, MO, USA. *e-mail: colin_donihue@fas.harvard.edu

© 2018 Springer Nature Limited. All rights reserved.

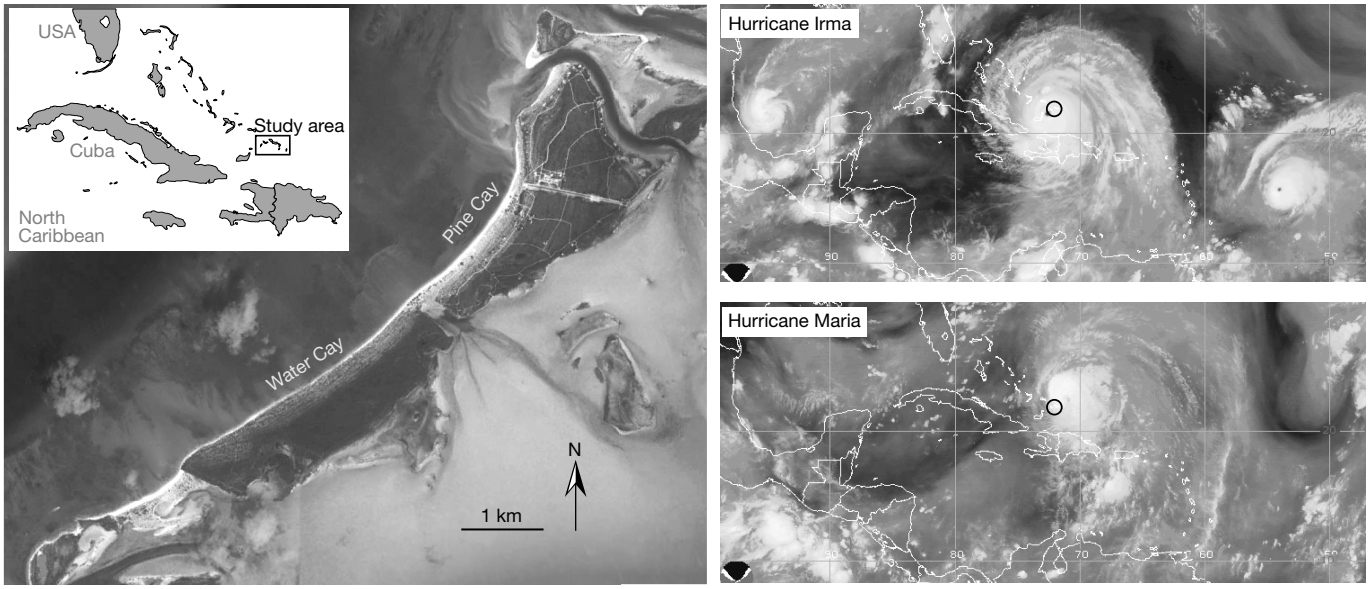


Fig. 1 | Location of Pine Cay and Water Cay with respect to Hurricanes Irma and Maria. Pine Cay and Water Cay are located in the Turks and Caicos Islands in the West Indies. They are home to the endemic Turks and Caicos anole, *Anolis scriptus*. On 8 September 2017, Hurricane Irma directly hit Turks and Caicos (black circle), shown in the water vapour satellite maps (from NOAA, www.goes.noaa.gov). Two weeks later, on 22 September 2017, Hurricane Maria struck Turks and Caicos. Map data: Google, © 2018 DigitalGlobe.

hurricanes. The variance in all six of these traits decreased among the surviving *A. scriptus* on Pine Cay, and decreased in four of the six on Water Cay (Supplementary Information), a result that is unlikely to have occurred by chance ($P = 0.019$ using the binomial test; analyses on principal component axes gave similar results, see Supplementary Information).

Second, survivors had traits associated with greater clinging ability. The positive relationship between toepad size and clinging capacity among anoles is well-established[17,19]; we confirmed this relationship for *A. scriptus* ($\beta \pm$ s.e.: 0.031 ± 0.01; $t_{86} = 3.26$; $P = 0.0016$; Methods). The larger toepads of surviving lizards support the hypothesis that natural selection favoured individuals with greater clinging capacity, which were able to withstand high winds. The observed shifts in limb length may have a similar functional explanation. Previous work has demonstrated that more force is needed to pull long-limbed anoles off a perch[18]. The longer forelimbs of surviving *A. scriptus* may have been beneficial during the hurricanes for this reason. However, the parallel decrease in femur and hind toe length was contrary to our predictions.

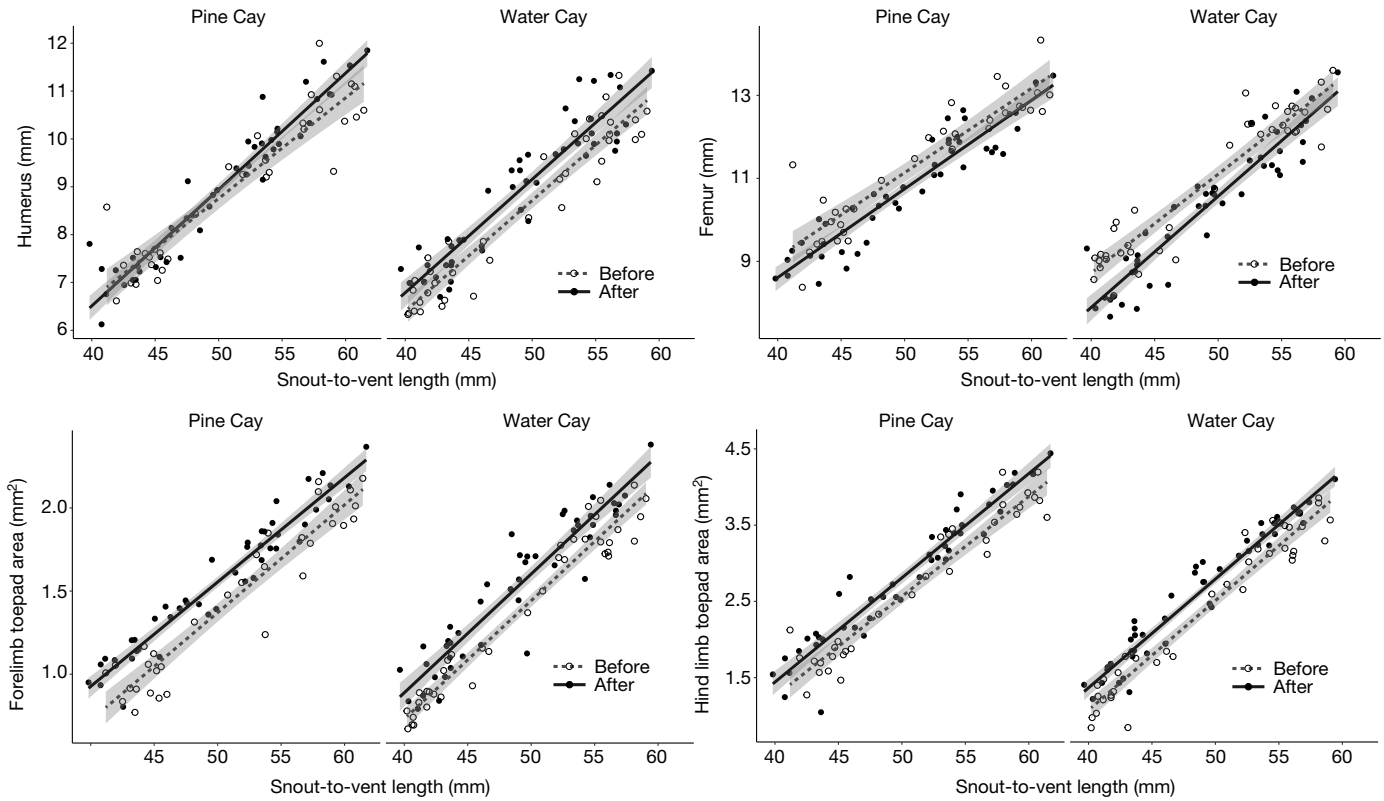


Fig. 2 | Parallel shifts in body-size-corrected humerus and femur length and surface area of the forelimb and hindlimb toepads for lizards on both Pine Cay and Water Cay. Dashed lines are the linear best fit of lizards measured before the hurricane ($n = 71$), which are represented by open circles. Filled circles are lizards from after the hurricane ($n = 93$) with solid lines of best fit. The grey-shaded areas correspond to 95% confidence intervals.

© 2018 Springer Nature Limited. All rights reserved.

Table 1 | Variance-standardized selection differentials

	Pine Cay	Water Cay
Snout-to-vent length	−0.17	−0.05
Humerus	0.40	0.8
Femur	−0.67	−0.89
Hindlimb longest toe	−0.67	−0.39
Forelimb toepad	1.18	1.04
Hindlimb toepad	0.88	1.08

See Methods for details. Variance-standardized selection differentials for humerus length, femur length, hindlimb longest toe length and toepad surface areas were calculated using body-size-corrected residuals.

Preliminary experimental trials (Supplementary Information) suggest one potential explanation: because of the posture *A. scriptus* adopts when exposed to high winds, longer hindlimbs could present a larger exposed surface area—which would increase the likelihood that a lizard would be blown off its perch.

Strong directional selection that favoured individuals able to hold tight during hurricane-force winds is a likely explanation for the parallel shifts in phenotype we observed. Of course, other explanations are possible: the hurricanes wrought environmental changes that affected vegetation structure, thermal microclimates and—most probably—food availability (though if lizards were starving, body condition would decrease, which on average it did not; $\beta \pm$ s.e.: 0.017 ± 0.019; $t_{161} = -0.867$; $P = 0.3874$; Supplementary Information). Although we cannot rule out these potential selective pressures, the fact that traits associated with clinging ability changed in ways that should increase survival in high winds lends credence to our hypothesis that the storms, and not their aftermath, drove these changes. Whatever the selective force was, selection must have been strong. For context, we calculated selection differentials for body size and the size-corrected morphological components that demonstrated a significant shift in their mean (Table 1), and found that they exceeded the majority of published selection differentials[20–22] and were of comparable magnitude to the selection experienced by Darwin's finches in two famously harsh periods[23].

Evolutionary processes other than natural selection could conceivably have produced these patterns, though we consider them implausible. We cannot rule out that lizard dispersal from elsewhere—unsampled microhabitats, other parts of the islands or other islands altogether—caused the observed phenotypic shifts. However, both Pine Cay and Water Cay are small islands with homogenous vegetation structure and all available microhabitats were thoroughly sampled. The appearance of morphologically different migrants would have increased rather than decreased trait variation in the recipient populations and would have had to occur in parallel on both islands (see Methods for additional sampling descriptions). We therefore discount this alternative explanation.

Attention to the evolutionary importance of phenotypic plasticity has increased in recent years, and adaptive phenotypic plasticity is more prevalent than previously realized[24]. We considered whether hurricane-induced phenotypic plasticity, either during the storms or as a result of post-hurricane conditions, could be responsible for the phenotypic shifts that we documented. In lizards, adult limb length can be affected by perch use during ontogeny[25,26], and bone shrinkage has been documented in starving marine iguanas[27]. By contrast, plasticity in toepad size has not been reported[28]. All of these studies, however, documented responses that occurred over long exposure periods or during ontogeny; our comparison spanned six weeks and was restricted to adults. Lastly, plastic decreases in skeletal elements have only previously been observed in response to food stress[27,29]; if food stress were the cause here—and our body condition data suggest this was not the case—it would have decreased the length of all skeletal elements. For these reasons, we consider phenotypic plasticity to be an unlikely explanation for the observed phenotypic shifts.

The role of extreme climate events in driving evolution is of pressing interest, and although some examples due to prolonged heat, cold or drought have been documented (reviewed in Grant et al.[11]), additional studies are needed in our era of rapidly changing and intensifying climate extremes[7,9]. Previous studies[9,11,30] have only alluded to the evolutionary ramifications of hurricanes. A 20-year study of a smaller lizard species, *Anolis sagrei*, on much smaller islands in the Bahamas—which are more vulnerable to hurricane-induced storm surges[16] than the islands in this study—showed that after the populations were hit by hurricanes, in the next sampling period mean tibia length had increased[30]. Whether these shifts were the result of natural selection and, if so, whether the selection had occurred during the hurricane itself or in the subsequent 8–48 months before the population was resampled remains unknown.

The long-term evolutionary consequences of these hurricanes on *A. scriptus* remain to be seen. Despite the extensive work on *Anolis* lizards, the biology of *A. scriptus* is little known. Indeed, our initial survey was conducted to gather baseline data on the natural history of the species in anticipation of a conservation project. That survey provided a serendipitous baseline from which to measure this selection event; future work is needed to determine whether the within-generational selection that we documented translates into evolutionary change across generations.

The macroevolutionary importance of infrequent but severe selective events such as these is an open question[11]. On the one hand, if such events are rare then natural selection in intervening periods might be expected to erase the signature of infrequent bouts of extreme selection. On the other hand, if hurricanes occur frequently enough—or if the selection is strong enough—then present-day populations may bear the mark of such events, in which case selection during normal years would not be able to fully explain current phenotypic distributions. In this light, it is notable that *Anolis* species occupying Caribbean islands have substantially larger toepads relative to their body size and habitat use than do congeners on mainland Central and South America[31]. Why this difference exists has long been questioned, but our findings suggest the possibility that hurricane-induced selection, a much more common occurrence on islands than on the mainland, may be responsible for this macroevolutionary pattern.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0352-3.

Received: 16 January 2018; Accepted: 11 June 2018;
Published online 25 July 2018.

1. Walker, L. R., Lodge, D. J., Brokaw, N. V. L. & Walde, R. B. An introduction to hurricanes in the Caribbean. *Biotropica* **23**, 313–316 (1991).
2. Scalley, T. H., Scatena, F. N., Lugo, A. E., Moya, S. & Estrada Rulz, C. R. Changes in structure, composition, and nutrients during 15 years of hurricane-induced succession in a subtropical wet forest in Puerto Rico. *Biotropica* **42**, 455–463 (2010).
3. Pavelka, M. S. M., McGoogan, K. & Steffens, T. S. Population size and characteristics of *Alouatta pigra* before and after a major hurricane. *Int. J. Primatol.* **28**, 919–929 (2007).
4. Willig, M. R. & Camilo, G. R. The effect of Hurricane Hugo on six invertebrate species in the Luquillo experimental forest of Puerto Rico. *Biotropica* **23**, 455–461 (1991).
5. Wiley, J. W. & Wunderle, J. M. The effects of hurricanes on birds, with special reference to Caribbean islands. *Bird Conserv. Int.* **3**, 319–349 (1993).
6. Endler, J. A. *Natural Selection in the Wild* (Princeton Univ. Press, Princeton, 1986).
7. Hegerl, G. C., Hanlon, H. & Beierkuhnlein, C. Climate science: elusive extremes. *Nat. Geosci.* **4**, 142–143 (2011).
8. Ummenhofer, C. C. & Meehl, G. A. Extreme weather and climate events with ecological relevance: a review. *Phil. Trans. R. Soc. B* **372**, 20160135 (2017).
9. Grant, P. R. Evolution, climate change, and extreme events. *Science* **357**, 451–452 (2017).
10. van de Pol, M., Jenouvrier, S., Cornelissen, J. H. C. & Visser, M. E. Behavioral, ecological and evolutionary responses to extreme climactic events: challenges and directions. *Phil. Trans. R. Soc. B* **372**, 20160134 (2017).

© 2018 Springer Nature Limited. All rights reserved.

11. Grant, P. R. et al. Evolution caused by extreme events. *Phil. Trans. R. Soc. B* **372**, 20160146 (2017).
12. Johnson, A. B. & Winker, K. Short-term hurricane impacts on a neotropical community of marked birds and implications for early-stage community resilience. *PLoS ONE* **5**, e15109 (2010).
13. Lodge, D. J. & McDowell, W. H. Summary of ecosystem-level effects of Caribbean hurricanes. *Biotropica* **23**, 373–378 (1991).
14. Batista, W. B. & Platt, W. J. Tree population responses to hurricane disturbance: syndromes in a south-eastern USA old-growth forest. *J. Ecol.* **91**, 197–212 (2003).
15. Wulff, J. L. Effects of a hurricane on survival and orientation of large erect coral reef sponges. *Coral Reefs* **14**, 55–61 (1995).
16. Spiller, D. A., Losos, J. B. & Schoener, T. W. Impact of a catastrophic hurricane on island populations. *Science* **281**, 695–697 (1998).
17. Crandell, K. E., Herrel, A., Sasa, M., Losos, J. B. & Autumn, K. Stick or grip? Co-evolution of adhesive toepads and claws in *Anolis* lizards. *Zoology* **117**, 363–369 (2014).
18. Kolbe, J. Effects of hind-limb length and perch diameter on clinging performance in *Anolis* lizards from the British Virgin Islands. *J. Herp.* **49**, 284–290 (2015).
19. Losos, J. *Lizards in an Evolutionary Tree: Ecology and Adaptive Radiation of Anoles* (Univ. of California Press, Berkeley, 2009).
20. Kingsolver, J. G. et al. The strength of phenotypic selection in natural populations. *Am. Nat.* **157**, 245–261 (2001).
21. Kingsolver, J. G. & Diamond, S. E. Phenotypic selection in natural populations: what limits directional selection? *Am. Nat.* **177**, 346–357 (2011).
22. Siepielski, A. M., DiBattista, J. D. & Carlson, S. M. It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecol. Lett.* **12**, 1261–1276 (2009).
23. Grant, P. R. & Grant, B. R. Evolution of character displacement in Darwin's finches. *Science* **313**, 224–226 (2006).
24. Forsman, A. Rethinking phenotypic plasticity and its consequences for individuals, populations and species. *Heredity* **115**, 276–284 (2015).
25. Losos, J. B. et al. Evolutionary implications of phenotypic plasticity in the hindlimb of the lizard *Anolis sagrei*. *Evolution* **54**, 301–305 (2000).
26. Kolbe, J. J. & Losos, J. B. Hind-limb length plasticity in *Anolis carolinensis*. *J. Herp.* **39**, 674–678 (2005).
27. Wikelski, M. & Thom, C. Marine iguanas shrink to survive El Niño. *Nature* **403**, 37–38 (2000).
28. Stuart, Y. E. et al. Rapid evolution of a native species following invasion by a congener. *Science* **346**, 463–466 (2014).
29. Lázaro, J., Dechmann, D. K. N., LaPoint, S., Wikelski, M. & Hertel, M. Profound reversible seasonal changes of individual skull size in a mammal. *Curr. Biol.* **27**, R1106–R1107 (2017).
30. Schoener, T. W., Kolbe, J. J., Leal, M., Losos, J. B. & Spiller, D. A. A multigenerational field experiment on eco-evolutionary dynamics of the influential lizard *Anolis sagrei*: a mid-term report. *Copeia* **105**, 543–549 (2017).
31. Macrini, T. E., Irschick, D. J. & Losos, J. B. Ecomorphological differences in toepad characteristics between mainland and island anoles. *J. Herp.* **37**, 52–58 (2003).

Acknowledgements This work was made possible thanks to the Pine Cay Homeowners Association and the US National Science Foundation. All procedures were approved by the Turks and Caicos DECR (Initial Permit: 17-08-02-14; Revisit Permit: 17-10-01-15) and Harvard University IACUC (26-11). We thank C. Santoro for help in the field and with the manuscript, and members of the Losos laboratory for valuable feedback. A.H. thanks the Caribaea Initiative for promoting science in the Caribbean and facilitating the networking that made the initial Turks and Caicos study possible.

Reviewer information *Nature* thanks E. Palkovacs and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.M.D., A.H., J.J.K. and J.B.L. conceived the study; C.M.D., A.H. and A.-C.F. conducted fieldwork; C.M.D., A.J.G., A.-C.F. and A.K. conducted analyses; C.M.D. and J.B.L. wrote the initial draft of the manuscript; and T.W.S., J.J.K., A.H., A.-C.F., A.J.G. and A.K. all contributed to interpretation of the data and to editing of subsequent drafts of the manuscript.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0352-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.M.D.
Competing interests: The authors declare no competing interests.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size.

Pine Cay and Water Cay are islands in the Turks and Caicos archipelago. Pine Cay, the larger of the two, is approximately 350 ha, sparsely inhabited by approximately 20 private residences, and covered with low vegetation (1–3 m in height). Water Cay (250 ha) was a separate island until the mid-1990s, when Water and Pine Cays were connected by a sand bridge deposited by a hurricane storm surge. Although some patches of vegetation grow between the islands, the narrow sand bar probably strongly inhibits the movement of lizards between the two islands; no lizards were observed on our trips crossing it.

A. scriptus, the Turks and Caicos anole, is a small (adult SVL 40–70 mm) and largely insectivorous lizard. It is abundant on Pine and Water Cays, readily detectable and typically found perched on vegetation at or below 3 m in height[32]. Our initial pre-hurricane survey of the anoles of these cays for an ongoing conservation project started on 28 August 2017 and lasted until 4 September 2017. As we did not anticipate investigating hurricane-induced selection, none of the captured lizards were permanently marked. Hurricane Irma hit Turks and Caicos on 8 September 2017, followed by Hurricane Maria on 22 September 2017. Our post-hurricane visit spanned from 16 October 2017 to 20 October 2017. To sample the lizard populations, C.M.D., A.-C.F. and A.H. walked a transect approximately 2-km long on each island, catching lizards whenever sighted with a noose and pole. These transect paths were repeated by C.M.D. and A.H. in the post-hurricane revisit. In total, we caught 71 adult lizards in the initial survey and 93 during the revisit. The increased number of lizards in the second visit reflects an increase in sampling time; care was taken to sample in all of the same microhabitats before and after the hurricane.

Both Pine Cay and Water Cay are small islands with homogeneous vegetation structure; consequently, microgeographic morphological variation within each island is likely to be minimal and we attempted to survey what variation does exist within our 2-km-long transects. We therefore think it unlikely that dispersal of individuals from elsewhere on the island into our sampling area would shift the population mean. Further, if migrants were morphologically different from the initial study population, it is unlikely that migration from outside the study site would shift phenotypic distributions in the same way on both islands. Moreover, if migrants arrived from one or more phenotypically distinct populations, the result would be more likely to increase than decrease the variation within the recipient population, in contrast to our findings. This last point also applies to the possibility that lizards were blown in from another island (the closest point given the direction of storm being North Caicos Island, 7 km away), an explanation that we consider to be even less likely.

An additional consideration inextricably linked to capturing samples of animal populations is that during our pre-hurricane surveys we failed to detect morphologically distinct lizards occupying unique microhabitats (for example, tree canopies). If those lizards became more apparent after the hurricanes, they could change the sample trait means, and spuriously suggest population-wide trait shifts. This possibility seems unlikely for several reasons: we thoroughly sampled all available microhabitats along the capture transects; *A. scriptus* individuals are rarely found above 3 m[32] and spend the majority of their time within 1.5 m of the ground where they are particularly easy to spot. More generally, previous studies have found no evidence to indicate that, within anole populations, individuals that use different microhabitats differ in limb or toepad characteristics[33,34]. The proportion of individuals with extreme values for the morphological traits in the post-hurricane samples suggests that mortality was probably very high; however, because we did not estimate population size or individually mark lizards, change in population size could not be estimated.

During both surveys, A.H. measured the morphology (snout-to-vent length, and length of the humerus, radius, metacarpal, longest forelimb toe, femur, tibia, metatarsal and longest hindlimb toe[35]) of each individual using digital calipers (Mitutoyo 500-752). In addition, a photograph was taken of the right fore and hind feet of each lizard, unless a digit was missing—in which case the left was photographed. All photographs were captured with an iPhone 7 using a Moment Macro Lens attachment. Using ImageJ (v.1.51a., W. Rasband, National Institute of Health, Bethesda), C.M.D. measured the toepad area of the longest toe three times for each forelimb and hindlimb. The repeated measurements were highly consistent and the average was used for analyses (intraclass correlation coefficient for forelimb toepad area was 98.9% (95% confidence interval (CI): 98.6–99.2%) and for hindlimb toepad area was 99.6% (95% CI: 99.5–99.7%). C.M.D. also counted the number of lamellae on the toepad of the longest toe on the right forelimb and hindlimb.

Lizard clinging performance measurements were taken using a Vernier Dual Range Force Sensor DFS-BTA with an acetate transparency as a gripping surface[17]. Following standard protocols[17], the right forelimb of each lizard was pulled down the surface by C.M.D. and the maximum force exerted by the forelimb toepads was recorded and analysed by A.-C.F. Methods for the wind speed behavioural study are presented in the Supplementary Information. All procedures were approved by Harvard IACUC (26-11) and all lizards were released unharmed after the experiment to their point of capture.

Analyses. All analyses were conducted in R[36] on adult lizards (female SVL >40 mm, male SVL >45 mm). We used a multivariate analysis of covariance to evaluate whether limb morphology varied between lizards on both islands before and after the hurricane. To do so, we log-transformed all morphometric measurements and used sex as a fixed effect and body size (SVL) as a covariate. The fixed effects of interest were island of origin (Water Cay or Pine Cay), hurricane treatment (pre-/post-) and their interaction. We did not detect a three-way interaction with sex. Following the multivariate analysis of covariance, we conducted post hoc general linear models on individual limb traits and SVL. An interaction of hurricane and island of origin was tested in the linear model and removed if the interaction was not significant. For all models, sex was included as a fixed effect. For all relative morphometric analyses (that is, comparisons taking into account differences in body size) we used SVL as a covariate. We used the 'lsmeans' and 'effects' packages in R to test for significance in the model estimates (see Supplementary Information for additional analysis details).

To calculate selection differentials, we transformed SVL as well as the body-size-corrected humerus, femur and toepad areas to have a zero mean and unit variance. The post-hurricane population means were then subtracted from the pre-hurricane population means to calculate the differential[37]. We compared these differentials to other published studies[20–22], most of which, however, were calculated over longer periods. We calculated body condition as residual of a linear regression between body mass and snout-to-vent length.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All morphological and performance data collected for this study are available on Dryad, https://doi.org/10.5061/dryad.2t41r64.

32. Laska, A. L. The structural niche of *Anolis scriptus* on Inagua. *Breviora* **349**, 1–6 (1970).
33. Kamath, A. & Losos, J. B. Does ecological specialization transcend scale? Habitat partitioning among individuals and species of *Anolis* lizards. *Evolution* **71**, 541–549 (2017).
34. Spendelow, P. *Niche Width and Morphological Variation (Anole)*. PhD thesis, Univ. of Washington (1985).
35. Irschick, D. J., Vanhooydonck, B., Herrel, A. & Meyers, J. Intraspecific correlations among morphology, performance and habitat use within a green anole lizard (*Anolis carolinensis*) population. *Biol. J. Linn. Soc.* **85**, 211–221 (2005).
36. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, 2017).
37. Lande, R. & Arnold, S. J. The measurement of selection on correlated characters. *Evolution* **37**, 1210–1226 (1983).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0359-9

Ecosystem restructuring along the Great Barrier Reef following mass coral bleaching

Rick D. Stuart-Smith[1]*, Christopher J. Brown[2], Daniela M. Ceccarelli[3] & Graham J. Edgar[1]

Global warming is markedly changing diverse coral reef ecosystems through an increasing frequency and magnitude of mass bleaching events[1–3]. How local impacts scale up across affected regions depends on numerous factors, including patchiness in coral mortality, metabolic effects of extreme temperatures on populations of reef-dwelling species[4] and interactions between taxa. Here we use data from before and after the 2016 mass bleaching event to evaluate ecological changes in corals, algae, fishes and mobile invertebrates at 186 sites along the full latitudinal span of the Great Barrier Reef and western Coral Sea. One year after the bleaching event, reductions in live coral cover of up to 51% were observed on surveyed reefs that experienced extreme temperatures; however, regional patterns of coral mortality were patchy. Consistent declines in coral-feeding fishes were evident at the most heavily affected reefs, whereas few other short-term responses of reef fishes and invertebrates could be attributed directly to changes in coral cover. Nevertheless, substantial region-wide ecological changes occurred that were mostly independent of coral loss, and instead appeared to be linked directly to sea temperatures. Community-wide trophic restructuring was evident, with weakening of strong pre-existing latitudinal gradients in the diversity of fishes, invertebrates and their functional groups. In particular, fishes that scrape algae from reef surfaces, which are considered to be important for recovery after bleaching[2], declined on northern reefs, whereas other herbivorous groups increased on southern reefs. The full impact of the 2016 bleaching event may not be realized until dead corals erode during the next decade[5,6]. However, our short-term observations suggest that the recovery processes, and the ultimate scale of impact, are affected by functional changes in communities, which in turn depend on the thermal affinities of local reef-associated fauna. Such changes will vary geographically, and may be particularly acute at locations where many fishes and invertebrates are close to their thermal distribution limits[7].

The 2016 mass bleaching event affected coral reefs world-wide, with catastrophic impacts reported in the Red Sea, central Indian Ocean, across the Pacific Ocean and in the Caribbean[3,8,9]. The Australian Great Barrier Reef (GBR), the largest coral reef system in the world, experienced the warmest temperatures on record for the region. An estimated 91.1% of reefs along the GBR experienced some bleaching[3], resulting in an estimated loss of approximately 30% of live coral cover over the following six months[10]. The event was thus comparable to the 1998 mass bleaching event in the Indian Ocean in terms of reported impacts on corals[2,11]. We surveyed 186 reef sites along the GBR and at less-studied isolated reefs in the Coral Sea before and after the 2016 bleaching event, and here we report reef- and regional-scale effects of the extreme thermal anomaly and loss of coral cover on the rich reef-associated fish and mobile invertebrate fauna. At each site, globally standardized Reef Life Survey census methods[12] were used to quantify changes to coral cover, reef fishes and mobile macroinvertebrates at multiple depths (overall mean, 6.7 m; range, 0.8–17.0 m). 'Before' data were obtained between 2010 and 2015, and 'after' data were obtained 8–12 months after bleaching.

As reported elsewhere[10], decreases in live hard coral cover were widespread (Fig. 1), although we found that the regional pattern was more spatially heterogeneous than previously described, when field surveys were standardized amongst shallow reef crest habitat[10]. Forty-four of the 186 surveyed sites experienced absolute declines in live coral cover that exceeded 10% (up to 51% loss for one site at Osprey Reef), with the northern Coral Sea reefs suffering the most consistent losses (Fig. 1a, b). The magnitude of coral-cover change was related to the local sea temperature anomalies (Fig. 1d and Extended Data Fig. 1), but coral loss varied considerably, and not all reefs in regions that experienced the greatest temperature anomalies experienced losses in live coral cover. In some cases, such as the central Coral Sea reefs, a history of cyclone damage meant that there was relatively little coral to lose. Thus, geographical patterns in pre-bleaching cover had a critical role in the realized effects of bleaching on corals (Fig. 1d). Coral-cover losses of the greatest magnitude occurred in disparate locations, including in the northern Coral Sea (Boot and Osprey Reefs; mean, 15% absolute cover loss, or approximately 40% of the pre-bleaching live coral cover), and the southern GBR (most southerly Swain Reefs; 28% loss, or 100% of pre-bleaching cover). The northern reefs in the GBR experienced the most extensive bleaching of those surveyed during the 2016 event[3], but not all of the reefs in that area suffered the extreme rates of live coral-cover loss that were observed more generally[10] (Fig. 1a, b). The fate of bleached corals can vary considerably[13,14], and a reasonable proportion of corals on some of these reefs must have regained their zooxanthellae and survived the bleaching event. Algal cover substantially increased across the majority of reefs that experienced coral declines (Fig. 1c and Extended Data Fig. 2).

Not all coral declines that were observed during the study could be assumed to be solely due to the bleaching event (other disturbances, such as cyclones, may have also had impacts on corals at particular locations; see Methods). To investigate the effects on reef fauna that could be most clearly attributable to the bleaching event, we quantified changes on a subset of reefs that experienced extreme heating and substantial live coral-cover loss (see Methods for criteria). On these reefs, the abundance of coral-eating fishes (corallivores) consistently declined, and declines in local fish species richness were also common (Extended Data Fig. 3). Such changes have previously been observed as rapid responses to coral bleaching events[5,15,16], and are clearly a concerning form of reef-scale biodiversity loss. These changes were not observed on a subset of comparison reefs that also experienced extreme heating, but that did not experience an observable loss of live coral cover (Extended Data Fig. 3). Other previously reported short-term effects of bleaching, such as increased herbivore abundance[15] in response to a boom in algal resources[5,16], occurred on some study reefs, but were not consistent features of those reefs with the clearest impacts on coral cover attributable to bleaching (Extended Data Fig. 3).

[1]Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania, Australia. [2]Australian Rivers Institute, Griffith University, Nathan, Queensland, Australia. [3]ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland, Australia. *e-mail: rstuarts@utas.edu.au

© 2018 Springer Nature Limited. All rights reserved.

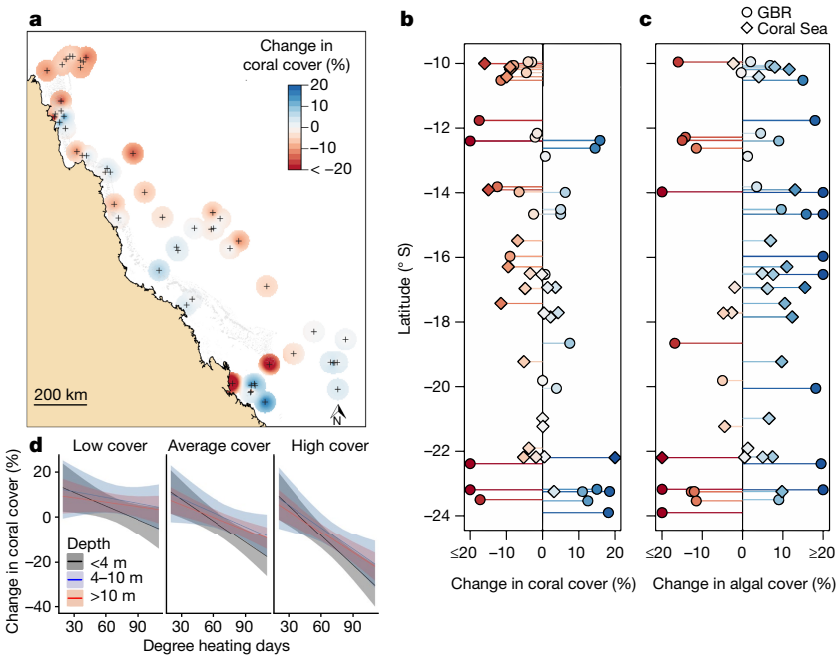


Fig. 1 | Observed changes in live hard coral cover from the 2016 mass bleaching event across the Coral Sea and GBR. a–c, Reefs in the Coral Sea showed relatively consistent losses of live corals (**a**, **b**) and gains in algal cover (**c**) in the north, whereas changes along the GBR were highly patchy. Absolute changes in live coral cover are mapped for individual sites ($n = 186$), with aggregation of sites at the reef scale shown as plusses ($n = 53$). **d**, Coral-cover loss was related to the local heat anomaly from January to March 2016 regardless of depth, an effect that increased in strength according to pre-bleaching cover of live corals. Average pre-bleaching cover for the region was 26% (middle), whereas low and high (left and right) are shown for ±1 s.d. (19%) from average pre-bleaching live coral cover. Effects in **d** are from Bayesian mixed-effects models, with shading representing 95% credible intervals.

Coherent patterns of ecological change were evident when assessing regional-scale trends between survey periods across the full range of sites surveyed. The latitudinal gradient in local species richness of mobile fauna[17] declined in slope through a combination of decreased local fish richness on northern reefs and markedly increased richness of macroinvertebrates and small cryptic fishes on southern reefs (Fig. 2 and Extended Data Fig. 4). The structure of fish communities on southern reefs became more similar to those in the north (Extended Data Fig. 5), a broad-scale homogenization that resulted in a slight decline in the overall number of fish species recorded across all surveys (from 532 to 494). Invertebrate communities also changed considerably between survey periods (Fig. 2 and Extended Data Fig. 4). This was characterized most clearly by sea urchins being found less frequently on northern reefs and in increased abundance on southern reefs after the bleaching event.

A key outcome of these changes was the regional alteration to the functional structure of reef communities, with potentially important consequences for the recovery of affected reefs. Functional richness (represented by the number of unique functional trait combinations comprised by fishes and invertebrates on each survey) increased on southern reefs, where the potential for local herbivory also increased through herbivorous fish biomass gains (Fig. 2 and Extended Data Figs. 4, 6) and patchy gains in the abundance of sea urchins. By contrast, the frequency of occurrence and biomass of fishes that scrape algae and microscopic autotrophs off coral–rock surfaces (scraping herbivores), and the frequency of sea urchins declined on northern reefs, whereas the biomass of plankton-feeding fishes increased (Figs. 2, 3 and Extended Data Fig. 4).

Most of these rapid, regional-scale ecological changes could not be linked to coral loss (Extended Data Fig. 4), and so cannot be assumed to be indirect effects of the bleaching event (or any other causes of coral degradation during the study). Some of these changes could nevertheless result from changes in the local composition and community structure of corals and algae, independently of the total amount of coral loss, but the spatial footprint of changes in the fishes and the invertebrates suggests that at least some of these changes were independent of changes in habitat. The consistency of ecological change along the latitudinal gradient differs from the heterogeneous patterns of the changes in coral and algal cover, particularly along the GBR, whereas the southern Coral Sea reefs showed very clear ecological changes, despite largely escaping bleaching. The loss of large predatory fishes in remote locations, such as in the northern GBR and on some reefs in the southern Coral Sea (Extended Data Fig. 6), could potentially be associated with expansion of the fishing footprint, but this needs further investigation. Changes in fishing pressure are unlikely to have resulted in most of the other coherent regional scale patterns of changes in the communities, because few herbivorous fishes, cryptic fishes and reef-dwelling invertebrates are targeted by fishers in this region.

Another potential explanation for the rapid restructuring of communities relates to more direct effects of region-wide anomalously warm temperatures and altered currents on the local occupancy patterns and abundance of different species[18]. Marine heat waves and short-term temperature variations have been shown to markedly affect temperate rocky reef communities[19], but have not been well-investigated on coral reefs. The sea temperatures that were experienced during the bleaching event (up to 32 °C in the northern GBR[18]) exceeded those at the warm limits of the distributions for the majority of reef fishes that were recorded in the region[7], and many species on northern reefs probably experienced thermal stress.

We used species temperature index (STI) values for fish species recorded during surveys to investigate the possibility that reduced species richness and altered trophic structure on the warmer northern reefs were due to disproportionate effects on species with an affinity for relatively cooler seas. STI values are derived from the realized thermal distributions of species across their entire range[7,20], and provide a nuanced and continuous measure of the ocean climate on which the distribution of each species is centred. On average, patterns of change in the frequency of occurrence of species in each trophic group were positively related to their STI values (Fig. 3). Specifically, those species that declined between surveys across northern reefs tended to be corallivores and scraping

© 2018 Springer Nature Limited. All rights reserved.

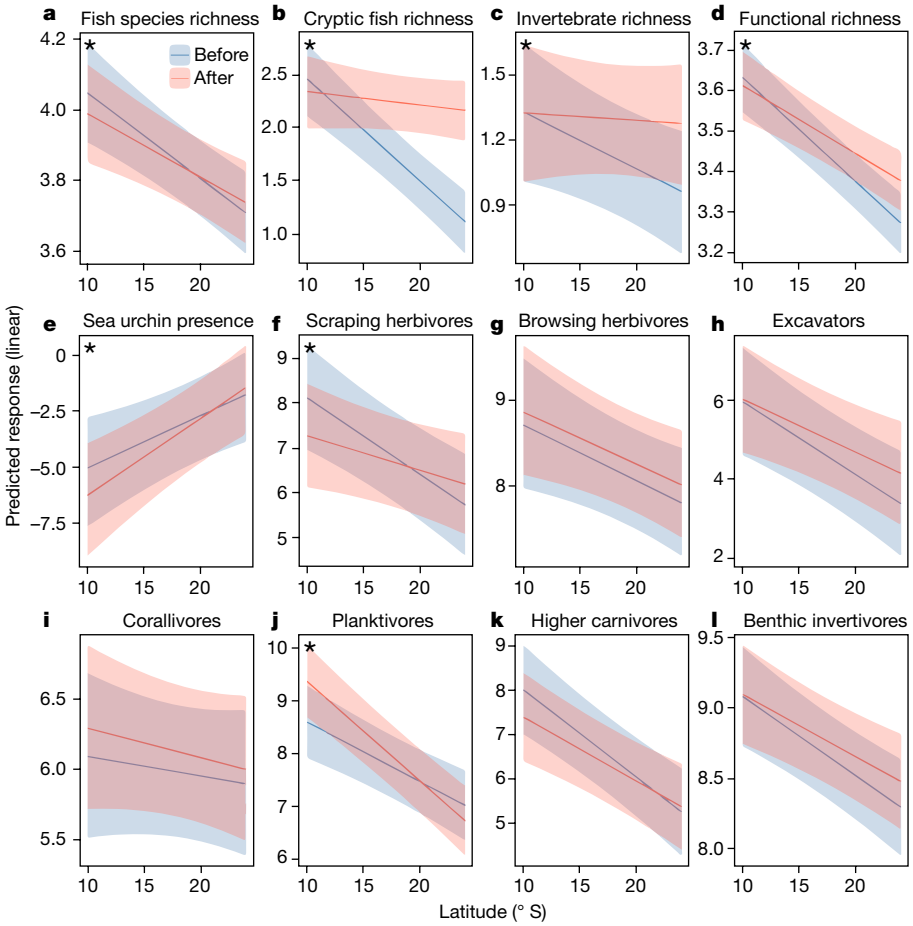


Fig. 2 | Changes in latitudinal trends in reef fish and invertebrate communities associated with the 2016 mass coral bleaching event. **a**–**d**, Plots show local richness (log scale). **e**, Plot shows sea urchin presence (log-odds). **f**–**l**, Plots show biomass in g per 500 m^2 (log scale). Latitudinal trends are median effect estimates from Bayesian generalized linear mixed-effects models ($n = 233$ site-by-depth-category combinations). Shaded regions show the marginal 95% credible intervals, and asterisks indicate those metrics for which the term for the change in latitudinal slopes (the interaction between latitudinal and time period effects) has 95% credible intervals that do not overlap zero (model effect sizes with credible intervals for all predictors are shown in Extended Data Fig. 4). *y* axes are on the link scale (log for Poisson and normal, and logit for sea urchins).

herbivores with distributions in relatively cooler waters; a pattern that was consistent for fish communities along both the GBR and Coral Sea, which have different biogeographical affinities[21] (Extended Data Fig. 5). Fishes that feed by excavating the coral–rock surface tended to have the warmest affinities (that is, higher STI values), and became more common in surveys in both the north and south (Fig. 3), although the increased frequency in the north did not translate to an increase in local biomass (Fig. 2 and Extended Data Fig. 4).

A bias in thermal affinities of reef fishes related to their trophic group has not previously been investigated in detail, and the generality of this phenomenon is unknown. In this case, the pattern was characterized by high variability (Fig. 3), and becomes increasingly influenced by excavators at the scale of the full GBR. The opposite situation may occur on temperate reefs, where herbivorous fishes have warmer STIs than other trophic groups[22]. Further investigation is needed to determine whether biases in STIs of trophic groups are idiosyncratic and location-dependent, or whether coherent geographical patterns emerge for particular trophic groups. The decreased frequency of occurrence of corallivores at northern reefs could also be related to coral mortality, with this effect potentially confounded by inferred effects of thermal stress (or other causes that were not investigated).

Ecological change on southern reefs included an increasing similarity of fish community structure to that on northern reefs (Extended Data Fig. 5), which is consistent with a potential influence of warmer temperatures, but could also result from altered currents and possible enhanced fish recruitment of northern species in the south. No clear signal of an influx of northern species recruits was evident, however, as the local richness of juveniles was no greater after the bleaching event than before (Extended Data Fig. 7). Instead, the majority of positive changes in the south related to taxa of relatively small adult body size—both invertebrates and cryptic fishes. These could be more sensitive to temperature changes and/or capable of increases in local population size more rapidly and/or could experience rapid numerical or behavioural release if predation pressure was reduced. Although less probable, release from predation may have resulted from minor decreases in the frequency of predatory fishes in the Coral Sea (Extended Data Fig. 6) and benthic invertebrate consumers in the GBR (Fig. 3).

Our broad-scale field surveys did not allow a definitive test of causation for the rapid regional ecological changes that were observed. Regardless of the causes, however, a critical feature is that the short-term effects of the bleaching may have been masked in some cases. For example, we observed an increase in fish species richness on a reef in the Swains area, despite concurrent coral devastation (albeit highly localized in a region that was otherwise little affected by bleaching[3]). Likewise, at the regional scale, local fish species richness increased on 40% of the surveyed reefs, despite mass bleaching, net coral loss and an overall decline in regional species richness. Such trends appear remarkable, given that a reduction in fish species richness has been amongst the most consistently and rapidly observed local ecological responses to coral loss observed in previous studies[6,16].

The observed regional-scale reshuffling and trophic reorganization appear to be extremely rapid, observable less than one year after

© 2018 Springer Nature Limited. All rights reserved.

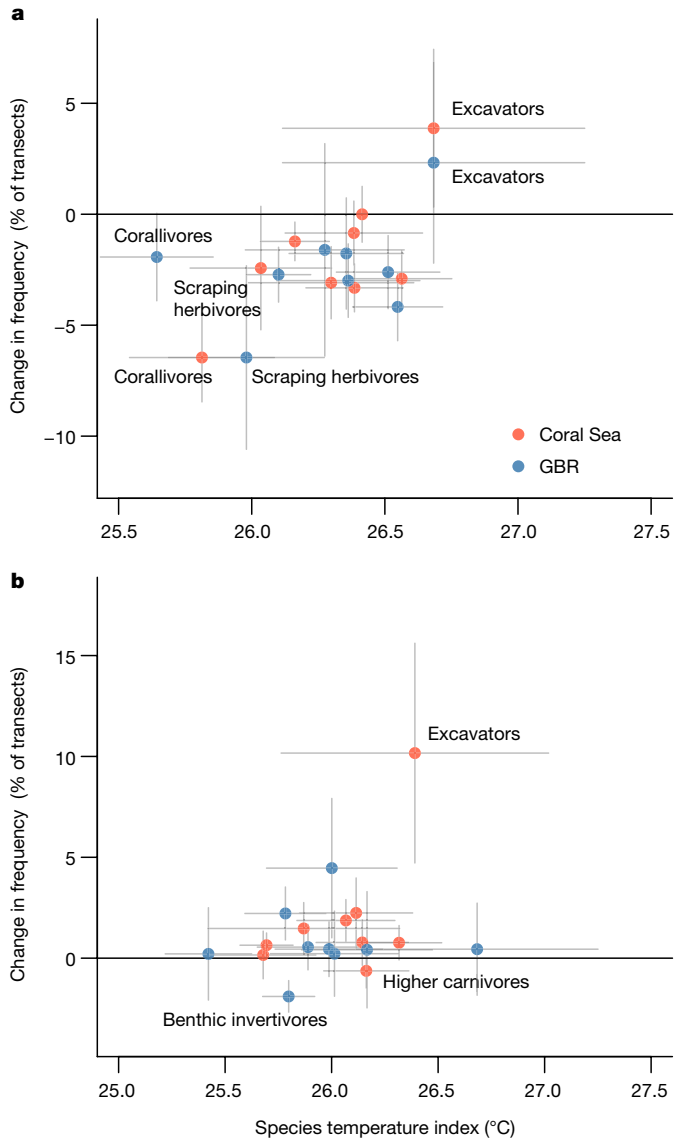


Fig. 3 | Changes in the frequency of fish functional groups with differing temperature affinities on reef surveys following the bleaching event. a, Reefs north of 12° S latitude. Corallivores and scraping herbivores, the coolest-affinity trophic groups (lower STI values, on average), declined in frequency, while excavators (higher STI values, on average) increased in frequency on transects at northern reefs. $n = 321$ species in GBR, 301 species in Coral Sea. **b**, Reefs south of 19° S latitude. Excavators became more common on transects at southern reefs in the Coral Sea. $n = 320$ species in GBR, 305 species in Coral Sea. Points are means of species in each trophic group, shown separately for species recorded in the GBR (blue) and Coral Sea (orange). Data are mean ± s.e.m.

the bleaching event. Rapid changes have previously been noted, such as increasing densities of herbivores[15], and have been hypothesized to be due to redistribution on reefs[23] rather than to a demographic response[5]. A substantial proportion of pre-bleaching surveys were undertaken in 2013 (ranging from 2010 to 2015), and many of the observed changes could have resulted from a number of consecutive warm years, rather than the single 2016 bleaching event. In addition to the 2016 event, the study period included two of the next nine warmest years on record for the GBR region (http://www.bom.gov.au/climate/change/; accessed September 2017). The observed patterns may thus in part represent accumulated responses over multiple exceptionally warm years, and could provide valuable signs of the potential trajectory of ecosystem change for a warmer future with increasingly prevalent extreme events[24].

Our observations of ecological change over an extreme heating event, with ecosystem consequences that are at least in part independent of coral mortality, may help to explain a lack of consistency among responses to bleaching observed in prior studies. For example, variability has previously been noted in herbivore responses[25], despite relatively consistent increases in algal resources following coral death[23]. This response is of critical importance, as the biomass of herbivorous fishes can be highly influential in determining the recovery trajectories of bleached reefs[2]. Scraping herbivores are considered to be particularly important for supporting reef recovery[26], and this group declined on northern reefs in our study. Whether losses of scraping herbivorous fishes in the northern GBR and Coral Sea will affect recovery of some of the most impacted reefs in the region remains an important question.

Ecosystem impacts of coral loss are likely to increase during the next decade in the GBR and Coral Sea if widespread erosion of dead corals occurs[1,5,6,15]. The extent to which the 2016 mass bleaching event proves ecologically catastrophic remains uncertain, as does the sum of accumulated effects from multiple bleaching events (as highlighted elsewhere[3,24,27]). However, rapid local recovery may occur on some reefs[28]. Either way, the trajectories of bleached reefs will be greatly influenced by the new community structures that we observed during a critical stage of reef recovery, and are thus inextricably linked with warming-related reshuffling of reef communities.

Overall, our results highlight the need for managers and researchers to consider broad spatial and temporal responses to the marine heating events amongst fishes and other biota, beyond the more readily observable impacts on coral habitat[29]. For example, potential ecological consequences of the changes that were observed in the northern GBR and Coral Sea could be exacerbated if herbivorous fishes were targeted by fisheries in these regions, whereas equivalent herbivore exploitation may not be an urgent management concern in locations where gains in herbivores occur (such as the southern GBR in our study). Likewise, functional changes in fish and invertebrate communities driven by extreme events may either complement or work against efforts to save reefs through restoration and assisted evolution of corals. Geographical location has been recognized as an important input into conservation planning and management from the perspective of considering patterns in ocean thermal regimes[30]. Our study highlights how location can additionally be important from the perspective of thermal affinities of community members. Accounting for the realized thermal niches of species in key functional groups may allow managers to more explicitly consider the trade-off between managing areas in which more species and functional groups are vulnerable to warming events, versus those in which fewer negative effects are expected. The former could potentially prolong local persistence of species and ecological stability by removing extractive pressures, and the latter may provide important reference areas for determining the importance of novel ecological interactions in shaping future reef ecosystems.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0359-9.

Received: 24 October 2017; Accepted: 22 May 2018;
Published online 25 July 2018.

1. Graham, N. A. J. et al. Lag effects in the impacts of mass coral bleaching on coral reef fish, fisheries, and ecosystems. *Conserv. Biol.* **21**, 1291–1300 (2007).
2. Graham, N. A. J., Jennings, S., MacNeil, M. A., Mouillot, D. & Wilson, S. K. Predicting climate-driven regime shifts versus rebound potential in coral reefs. *Nature* **518**, 94–97 (2015).
3. Hughes, T. P. et al. Global warming and recurrent mass bleaching of corals. *Nature* **543**, 373–377 (2017).
4. Poloczanska, E. S. et al. Responses of marine organisms to climate change across oceans. *Front. Mar. Sci.* **3**, 62 (2016).
5. Pratchett, M. S. et al. Effects of climate-induced coral bleaching on coral-reef fishes—ecological and economic consequences. *Oceanogr. Mar. Biol. Annu. Rev.* **46**, 251–296 (2008).

© 2018 Springer Nature Limited. All rights reserved.

6. Graham, N. A. J. et al. Dynamic fragility of oceanic coral reef ecosystems. *Proc. Natl Acad. Sci. USA* **103**, 8425–8429 (2006).
7. Stuart-Smith, R. D., Edgar, G. J. & Bates, A. E. Thermal limits to the geographic distributions of shallow-water marine species. *Nat. Ecol. Evol.* **1**, 1846–1852 (2017).
8. Perry, C. T. & Morgan, K. M. Bleaching drives collapse in reef carbonate budgets and reef growth potential on southern Maldives reefs. *Sci. Rep.* **7**, 40581 (2017).
9. Morgan, K. M., Perry, C. T., Johnson, J. A. & Smithers, S. G. Nearshore turbid-zone corals exhibit high bleaching tolerance on the Great Barrier Reef following the 2016 ocean warming event. *Front. Mar. Sci.* **4**, 224 (2017).
10. Hughes, T. P. et al. Global warming transforms coral reef assemblages. *Nature* **556**, 492–496 (2018).
11. Goreau, T., McClanahan, T., Hayes, R. & Strong, A. Conservation of coral reefs after the 1998 global bleaching event. *Conserv. Biol.* **14**, 5–15 (2000).
12. Edgar, G. J. & Stuart-Smith, R. D. Systematic global assessment of reef fish communities by the Reef Life Survey program. *Sci. Data* **1**, 140007 (2014).
13. Hoegh-Guldberg, O. & Salvat, B. Periodic mass-bleaching and elevated sea temperatures: bleaching of outer reef slope communities in Moorea, French Polynesia. *Mar. Ecol. Prog. Ser.* **121**, 181–190 (1995).
14. Hoegh-Guldberg, O. Climate change, coral bleaching and the future of the world's coral reefs. *Mar. Freshw. Res.* **50**, 839–866 (1999).
15. Garpe, K. C., Yahya, S. A. S. & Lindahl, U. & Öhman, M. C. Long-term effects of the 1998 coral bleaching event on reef fish assemblages. *Mar. Ecol. Prog. Ser.* **315**, 237–247 (2006).
16. Wilson, S. K., Graham, N. A. J., Pratchett, M. S., Jones, G. P. & Polunin, N. V. C. Multiple disturbances and the global degradation of coral reefs: are reef fishes at risk or resilient? *Glob. Change Biol.* **12**, 2220–2234 (2006).
17. Edgar, G. J. et al. Abundance and local-scale processes contribute to multi-phyla gradients in global marine diversity. *Sci. Adv.* **3**, e1700419 (2017).
18. Wolanski, E., Andutta, F., Deleersnijder, E., Li, Y. & Thomas, C. J. The Gulf of Carpentaria heated Torres Strait and the Northern Great Barrier Reef during the 2016 mass coral bleaching event. *Estuar. Coast. Shelf Sci.* **194**, 172–181 (2017).
19. Bates, A. E., Stuart-Smith, R. D., Barrett, N. S. & Edgar, G. J. Biological interactions both facilitate and resist climate-related functional change in temperate reef communities. *Proc. R. Soc. B* **284**, 20170484 (2017).
20. Stuart-Smith, R. D., Edgar, G. J., Barrett, N. S., Kininmonth, S. J. & Bates, A. E. Thermal biases and vulnerability to warming in the world's marine fauna. *Nature* **528**, 88–92 (2015).
21. Edgar, G. J., Ceccarelli, D. M. & Stuart-Smith, R. D. *Assessment of Coral Reef Biodiversity in the Coral Sea*. Report for the Department of the Environment (Reef Life Survey Foundation, 2015).
22. Vergés, A. et al. Long-term empirical evidence of ocean warming leading to tropicalization of fish communities, increased herbivory, and loss of kelp. *Proc. Natl Acad. Sci. USA* **113**, 13791–13796 (2016).
23. Diaz-Pulido, G. & McCook, L. J. The fate of bleached corals: patterns and dynamics of algal recruitment. *Mar. Ecol. Prog. Ser.* **232**, 115–128 (2002).
24. Hughes, T. P. et al. Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. *Science* **359**, 80–83 (2018).
25. McClanahan, T., Maina, J. & Pet-Soede, L. Effects of the 1998 coral morality event on Kenyan coral reefs and fisheries. *Ambio* **31**, 543–550 (2002).
26. Mumby, P. J. The impact of exploiting grazers (Scaridae) on the dynamics of Caribbean coral reefs. *Ecol. Appl.* **16**, 747–769 (2006).
27. Hughes, T. P. et al. Coral reefs in the Anthropocene. *Nature* **546**, 82–90 (2017).
28. Ferrari, R. et al. Quantifying the response of structural complexity and community composition to environmental change in marine communities. *Glob. Change Biol.* **22**, 1965–1975 (2016).
29. Hoegh-Guldberg, O. & Bruno, J. F. The impact of climate change on the world's marine ecosystems. *Science* **328**, 1523–1528 (2010).
30. Chollett, I., Enríquez, S. & Mumby, P. J. Redefining thermal regimes to design reserves for coral reefs in the face of climate change. *PLoS ONE* **9**, e110634 (2014).

Acknowledgements We thank the Reef Life Survey (RLS) divers and boat skippers who assisted with field surveys, including D. and J. Shields, I. Donaldson and S. Griffiths, and A. Cooper, J. Berkhout and E. Clausius at the University of Tasmania for logistics and data management; J. Stuart-Smith, S. Baker, A. Bates and N. Barrett for further support in the development of RLS, fieldwork and concepts explored in the paper. Development of RLS was supported by the former Commonwealth Environment Research Facilities Program, and analyses were supported by the Marine Biodiversity Hub, a collaborative partnership supported through the Australian Government's National Environmental Science Programme (NESP), and by the Australian Research Council. Funding and support for the GBR and Coral Sea RLS field surveys was provided by The Ian Potter Foundation and Parks Australia. Permits were provided by Parks Australia and the Great Barrier Reef Marine Park Authority. C.J.B. was supported by a Discovery Early Career Researcher Award (DE160101207) from the Australian Research Council.

Reviewer information *Nature* thanks J. Bruno, R. Ferrari and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.J.E. and R.D.S.-S. collected the data with the assistance of other Reef Life Survey divers; C.J.B. undertook the data analysis and preparation of figures with assistance from R.D.S.-S.; D.M.C. analysed the photoquadrats for benthic cover data; R.D.S.-S. drafted the paper with input from all other authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0359-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0359-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.D.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Survey methods. Standardized data were obtained on fishes, mobile invertebrates, coral and algae along 768 50-m underwater transects by trained scientific and recreational divers who participated in the citizen science Reef Life Survey (RLS) program. Full details of census methods are provided elsewhere[12,31,32], and an online methods manual (https://reeflifesurvey.com) describes all data-collection methods. Data quality and training of divers have previously been described[12,33]. All observed fish species were counted in duplicate 5-m-wide transect blocks and aggregated as densities per 500-m^2 transect, and cryptic fishes and mobile invertebrates >2.5 cm total length in duplicate 1-m-wide transect blocks (aggregated to 100-m^2 transect area) on the same transect lines. Fish length and abundance estimates were converted to biomass using species-specific length–weight coefficients obtained from FishBase (https://fishbase.org), as used in previous studies with the RLS data[34,35]. Invertebrate classes used for this study were Asteroidea, Cephalopoda, Crinoidea, Echinoidea, Gastropoda, Holothuroidea and Malacostraca. All individuals from these classes exceeding 2.5 cm total length were included in richness estimates for invertebrates, and in functional richness analyses.

Photoquadrats were taken vertically downward of the substrate every 2.5 m along each of the same transect lines, and later scored using a grid overlay of 5 points per image, 100 points per transect. Categories of benthic cover scored were from a set of 50 morphological and functional groups of algae and corals (Extended Data Table 1), as previously described[32] and aligning with the standard Australian hierarchical benthic classification scheme[36]. Analyses undertaken for this study were based on the sum of all live hard coral categories (that is, percentage of live hard coral per transect), and the sum of all algal categories (percentage of algal cover per transect), with categories listed in Extended Data Table 1.

Survey design. Matching before–after bleaching surveys were undertaken at 186 GPS-referenced sites at 53 reefs (see Fig. 1 for distribution of reefs; mean = 3.5 sites per reef) along the full length of the GBR and western Coral Sea region within the Australian Exclusive Economic Zone. At each site, multiple surveys (mean = 2.1 transects per site) were undertaken at different depths, with transects laid along a depth contour. Depths were binned (see 'Covariates'), such that the site-by-depth bin was the level of replication, making 233 matching site-by-depth replicates surveyed both before and after the bleaching event.

Different divers often surveyed the fishes and the invertebrates along the same transect line. Pre-bleaching surveys were mostly undertaken from a survey cruise along the entire GBR and Coral Sea in 2015 (42% of pre-bleaching surveys) and a previous survey cruise through the GBR and Coral Sea in 2013 (39% of pre-bleaching surveys). Additional 'before data' (19%) were collected at Lizard Island, Great Keppel Island and the Whitsundays in 2010, and some sites in the central Coral Sea and GBR in 2012. All post-bleaching data were collected during a survey cruise through the entire region from November 2016 to March 2017. No strong biases were apparent in the interval between pre- and post-bleaching surveys along the latitudinal gradient or locations experiencing different heating anomalies (Extended Data Fig. 8).

Seven of the 10 divers who undertook pre-bleaching surveys also undertook post-bleaching surveys, and G.J.E. and R.D.S.-S. together undertook 45% of all fish surveys (and led 85% of survey voyages before and 92% after the bleaching event). There was thus a substantial element of consistency in divers during the study. To explore the effect of different divers undertaking surveys at different times, however, we reran the models for Fig. 2 and Extended Data Fig. 4 with 'diver' included as a random effect. This resulted in no changes in the effect sizes or conclusions. Therefore, results are presented for models without the diver effect, so that marginal uncertainty intervals include site-to-site variation but not observer variation.

Species traits. All fishes and invertebrates were allocated into one of the following trophic groups: corallivores, scraping herbivores, benthic invertivores, algal farmers, browsing herbivores, omnivores, planktivores, higher carnivores, excavators, detritivores, suspension feeders and cleaners. Additional traits used for calculation of functional richness were: maximum body size (included as 10-cm bins up to 50 cm, and all species which grow to >50 cm binned together), and water column position (benthic, demersal, pelagic site-attached and pelagic non-site-attached). All traits were taken from a previously published dataset[37]. Functional richness was calculated as the richness of functional entities per 50-m transect, in which all species with the same combination of trait levels for those three traits were considered functionally equivalent.

STI values were taken for each species from a previously published dataset[20], and represent the midpoint between the 5th and 95th percentile of local mean sea-surface temperature values from all occurrence locations of the species. It thus represents the centre of each species' range when expressed as a range of sea temperatures experienced across its distribution, and provides a nuanced means of ordering species by their preferences for warmer or cooler environments. Full details, including discussion of strengths and weaknesses, are provided in previous publications[7,20].

Covariates. The mean depth contour of each reef transect was recorded by divers during surveys, with surveys then allocated into three depth bins (<4 m, 4–10 m and >10 m). For any before–after comparisons, we first obtained the mean values of univariate responses taken from among all transects within each depth bin at a given site (that is, site-by-depth bin combinations). This gave 233 site-by-depth combinations, with a mean of 76 sites and 35.3 reefs per depth class. For each site, we also applied a four-level categorical measure for wave exposure: (1) sheltered, with only wind waves from non-prevailing direction; (2) wind-generated waves from the prevailing direction; (3) exposed to ocean swells, either indirectly with exposure to prevailing winds, or directly but sheltered from prevailing winds; or (4) exposed to open ocean swell from prevailing direction. There was a mean of 62 sites and 24 reefs per exposure category. Reef habitat categories are often used for ecological studies of coral reefs (for example, slope, crest, flat and lagoon), but delineation between similar or adjacent habitats can sometimes be difficult. Instead of making these delineations for our survey sites, we considered that these two environmental axes of wave exposure and depth together appropriately capture the important variation between such reef habitat classifications with respect to their importance in describing potential for bleaching[2].

Sea-surface temperature anomalies used in analyses relating coral-cover change to degree heating days (DHD) was obtained from the ReefTemp Next Generation[38]. Fine scale anomalies for the period of January–March 2016 were matched to survey sites.

Analysis of coral- and algal-cover change. We modelled the response of change in coral and algal cover as a function of DHD using a Bayesian mixed-effects model ($n = 211$ site-depth combinations where benthic-cover data were available). Additional fixed covariates included the depth of survey, the four wave exposure categories, a factor for whether the survey was in the GBR or the Coral Sea, the initial cover of corals or algae, an interaction between DHD and depth and an interaction between DHD and initial cover. We included a random effect for reef. We did not include a random effect for sites nested within reefs because only 36 had measurements at more than one depth across both time periods (before and after bleaching). Change in coral and algal cover was modelled with Gaussian errors and standard model checks confirmed that this assumption was appropriate. We scaled the variance of the model by the number of years between before and after surveys (maximum = 7 years, mean = 3.3 years), because we expect greater variance in the measured change in coral cover when those measurements were taken a longer time apart. We compared models with and without the variance scaling using the widely applicable information criteria (WAIC)[39,40]. The WAIC indicated that for the coral-cover model with variance scaling provided an enhanced fit to the data (1,658 versus 1,721), whereas for the algae-cover model the unscaled model had an enhanced fit (1,801 versus 1,806), so we present results from these best models. However, the estimated effects of the covariates were nearly identical regardless of model used in both cases.

We present the median estimated effects of DHD on coral and algal cover in Extended Data Fig. 1, and credible intervals are 95% quantiles. We also predict median change and the 95% marginal credible intervals for change in coral and algal cover across the range of DHD for each depth category values for low wave exposure reefs in the GBR (Fig. 1d). Credible intervals for predictions were integrated across all random effects, so they should be interpreted as effect sizes relative to variation across reefs. For the coral model, there was a strong interaction effect of initial coral cover with DHD, so we separately plotted predicted effects for the mean initial coral cover and ±1 s.d. in initial coral cover.

We fitted the Bayesian mixed-effects models using the INLA framework[41] implemented in the R programming language[42] using the INLA R package (version 17.06.20; http://r-inla.org, accessed 4 October 2017). The prior for the precision on the random effect used the log-gamma prior with shape = 1 and rate = 1×10^{-5}, although use of other standard priors did not change the results. Priors for fixed effects had mean = 0 and precision = 0.001.

Mapped coral change values in Fig. 1a represent absolute change in live hard coral cover at each site, with the change values interpolated using an inverse-distance weighting and a buffer of 50 km applied from around each reef surveyed (implemented with the gstat package in the R program[43]). Symbols on the map thus represent the locations of the reefs, although coral change values come from the aggregation of smaller scale data at individual sites within reefs.

Comparison of bleaching-impacted and unaffected reefs. To isolate ecological impacts most likely arising from bleaching-associated coral loss, we used the following criteria to define 'bleaching-impacted' reefs: (1) pre-bleaching live hard coral cover >20% on average (across all transects at the reef). This meant that the starting community was more likely one to be comprised of coral-associated fish and invertebrate species; (2) loss of live coral cover >40% of pre-bleaching values, on average; and (3) experienced more than 40 DHD. These criteria were collectively used as a means to show the maximum likely impact of the loss of coral

© 2018 Springer Nature Limited. All rights reserved.

from bleaching, by ensuring there was adequate coral cover to start with (criterion 1), and that coral losses were at least typical of the mortalities observed in other studies[11] (criterion 2), while providing some confidence that observed coral loss was most likely attributable to bleaching (criterion 3). We cannot be certain about the latter (see comments below about other potential impacts on coral during the study period), but 40 DHDs well exceeds the threshold for bleaching as identified previously[3] for this same bleaching event. Reefs defined as 'bleaching-impacted' were widely dispersed along the GBR and Coral Sea (Extended Data Fig. 3).

To provide an objective contrast with these reefs, we also selected reefs that were clearly unaffected by bleaching. We used the same criteria as above, but instead of losing at least 40% of live coral cover, we selected only those that experienced >1% mean gain in live hard corals on average (using the mean percentage of coral-cover change, rather than mean pre-bleaching minus mean post-bleaching cover). For all bleached and unaffected reefs, we examined responses in key metrics of the coral, fish and invertebrate communities, as shown in Extended Data Fig. 3.

Regional community-structure change. Non-metric multidimensional scaling was undertaken separately on reef fish and invertebrate community data to show broad regional change in community structure and visualize consistencies in the direction of community change among regions between before and after surveys (Extended Data Fig. 5). Mean biomass per fish species (kg per 500 m^2) and mean abundance per mobile invertebrate species (individuals per 100 m^2) were calculated separately across all surveys within each 2° latitudinal band for the GBR and Coral Sea. Biomass and abundance data were log-transformed and Bray–Curtis dissimilarity matrices used for ordination of each. The analysis was undertaken in PRIMER[44], with symbols subsequently colour-coded for data collected before and after the bleaching event, and labels to indicate GBR and Coral Sea regions.

Analysis of regional-scale ecological changes. We analysed the response of nine fish and invertebrate metrics to bleaching using Bayesian generalized linear mixed-effects models (GLMMs). Each metric value on each survey was modelled with covariates for latitude, depth, coral cover, protection status (no-take 'green' zone versus all other zone types), GBR or coral sea, wave exposure, time (before or after the bleaching event) and an interaction between time and latitude. The interaction was included to allow for the possibility that latitudinal gradients in each metric changed from before to after the bleaching event. We included random effects for reefs and sites within reefs.

We chose error distributions appropriate for each metric. These were: Poisson with a log-link for the richness metrics, log-normal with an identity link for the biomass metrics, and binomial with a logit link for urchin presence. We added 0.5 to the logged biomass data so that zero values were not excluded. Checks of residuals confirmed that a log-normal distribution was appropriate for the biomass data. Rootograms[45] and Dunn–Smyth residuals[46] were used to confirm the count models were fitted appropriately.

We used the INLA framework to fit the Bayesian GLMMs, using the same settings as for the coral change model. We give effect sizes as median effects of each covariate in Fig. 2 and Extended Data Fig. 4 with 95% credible intervals. Credible intervals that did not overlap zero in Extended Data Fig. 4 are identified by asterisks. We also predict metrics across the latitudinal gradient before and after the bleaching event with marginal 95% credible intervals. Predictions across latitude were made for a reef of <4-m depth, with the mean level of hard coral cover, inside a protected area with low wave exposure and for the GBR. Thus positive and negative effects in Extended Data Fig. 4 can be interpreted in relation to these levels of the relevant covariates. Choosing other covariate values for the baseline would affect the magnitude of the patterns but not the overall trend.

Possible recruitment events. We tested whether patterns in richness before and after bleaching could be related to a coincident fish recruitment event. We analysed the mean richness of juvenile fishes per reef (29 reefs in total from extreme north and south) as a function of three binary covariates: before versus after bleaching, Coral Sea versus GBR and north versus south, using a linear model, implemented in the INLA framework[41] from the R programming language[42]. Juveniles were defined as any individuals that were 10 cm or less, for species that exceed 12.5 cm in maximum size. No significant change in the richness of juveniles was evident before and after bleaching (mean difference = 1.70 with lower and upper 95% credible intervals of –0.6 and 4.0), with the distribution of data shown in Extended Data Fig. 7.

Other potential effects on results. Few trends in fishes and invertebrates were related to changes in coral cover when considered at the scale of the whole study region, and primary study conclusions do not rest on the assumption that all observed coral mortality was driven by the 2016 bleaching event. Cyclones, crown-of-thorns starfish, and pollution and sediment from riverine outputs are other potential impacts on corals across the region. We checked the database of past tropical cyclone tracks on the Bureau of Meteorology website (http://www.bom.gov.au/cyclone/history/index.shtml, accessed 7 April 2018) for intersection of cyclone tracks with our survey sites. Surveys were completed before cyclone Debbie (2017), and the only surveys done before cyclone Yasi (2011) (Lizard Island, Port Douglas, Whitsundays, Keppel) were in areas outside of the destructive path of this cyclone. However, cyclone Ita was reported to have impacts on corals in the Lizard Island area during the study period[47], and there is a possibility that other smaller cyclones caused highly localized impacts. Thus, caution is required in ruling out cyclone damage as contributing to coral-cover changes observed in some locations.

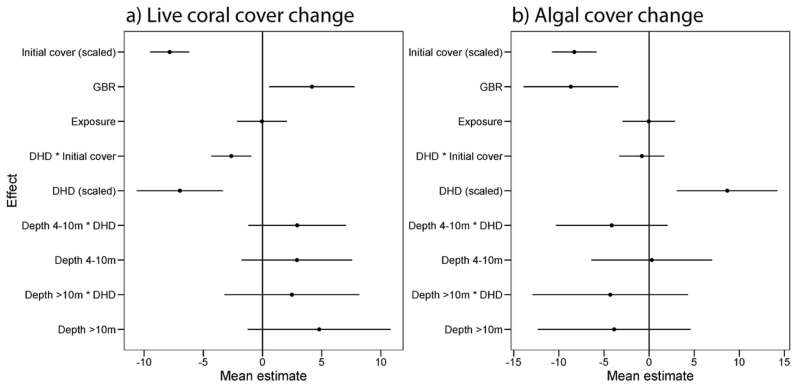
We cannot be certain that crown-of-thorns starfish did not affect coral cover at our sites in between surveys, but these are also recorded on the surveys of mobile invertebrates and were found in extremely low densities (mean = 1.4 individuals per 50 m^2 when found, only at 15 sites). It is not impossible that a wave of crown-of-thorns starfish came through and reduced live coral cover at a small number of sites, but such effects at this very small number of sites would unlikely have a detectable impact on results or conclusions of the study. Likewise, pollution and sediment from riverine sources could not have been responsible for any changes in the Coral Sea (>250 km offshore), and would be unlikely to have impacted any sites other than a small number of inshore locations. No substantial pollution events (for example, oil spills) were noted near survey locations in the period. Regardless, care is required in inferring causality for observed coral-cover change in this study, and no assumption should be made that all coral loss was attributable to bleaching.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Raw reef fish and invertebrate abundance data and photoquadrats of coral cover are available online through the Reef Life Survey website: https://reeflifesurvey.com.

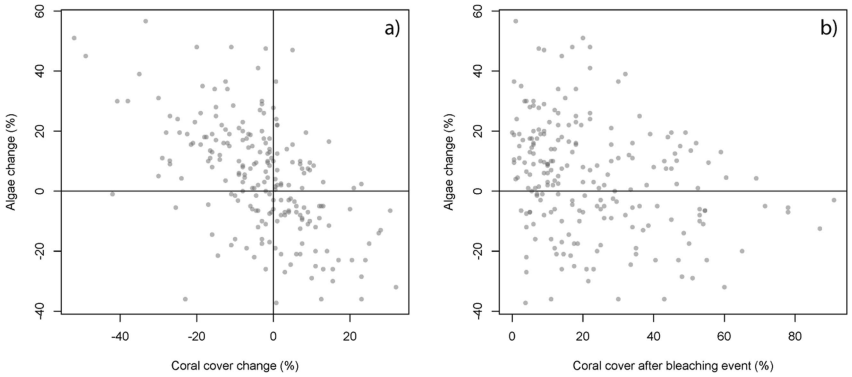
31. Edgar, G. J., Stuart-Smith, R. D., Cooper, A., Jacques, M. & Valentine, J. New opportunities for conservation of handfishes (Family Brachionichthyidae) and other inconspicuous and threatened marine species through citizen science. *Biol. Conserv.* **208**, 174–182 (2017).
32. Cresswell, A. K. et al. Translating local benthic community structure to national biogenic reef habitat types. *Glob. Ecol. Biogeogr.* **26**, 1112–1125 (2017).
33. Edgar, G. J. & Stuart-Smith, R. D. Ecological effects of marine protected areas on rocky reef communities—a continental-scale analysis. *Mar. Ecol. Prog. Ser.* **388**, 51–62 (2009).
34. Edgar, G. J. et al. Global conservation outcomes depend on marine protected areas with five key features. *Nature* **506**, 216–220 (2014).
35. Duffy, J. E., Lefcheck, J. S., Stuart-Smith, R. D., Navarrete, S. A. & Edgar, G. J. Biodiversity enhances reef fish biomass and resistance to climate change. *Proc. Natl Acad. Sci. USA* **113**, 6230–6235 (2016).
36. Althaus, F. et al. A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the CATAMI classification scheme. *PLoS ONE* **10**, e0141039 (2015).
37. Stuart-Smith, R. D. et al. Integrating abundance and functional traits reveals new global hotspots of fish diversity. *Nature* **501**, 539–542 (2013).
38. Garde, L., Spillman, C. M., Heron, S. & Beeden, R. ReefTemp next generation: a new operational system for monitoring reef thermal stress. *J. Oper. Oceanogr.* **7**, 21–33 (2014).
39. Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016 (2014).
40. Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**, 3571–3594 (2010).
41. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Stat. Soc. B* **71**, 319–392 (2009).
42. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, 2013).
43. Pebesma, E. J. Multivariable geostatistics in S: the gstat package. *Comput. Geosci.* **30**, 683–691 (2004).
44. Anderson, M. J., Gorley, R. N. & Clarke, K. R. PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods (PRIMER-E, Plymouth, 2008).
45. Kleiber, C. & Zeileis, A. Visualizing count data regressions using rootograms. *Am. Stat.* **70**, 296–303 (2016).
46. Dunn, P. K. & Smyth, G. K. Randomized quantile residuals. *J. Comput. Graph. Stat.* **5**, 236–244 (1996).
47. Cheal, A. J., MacNeil, M. A., Emslie, M. J. & Sweatman, H. The threat to coral reefs from more intense cyclones under climate change. *Glob. Change Biol.* **23**, 1511–1524 (2017).

© 2018 Springer Nature Limited. All rights reserved.



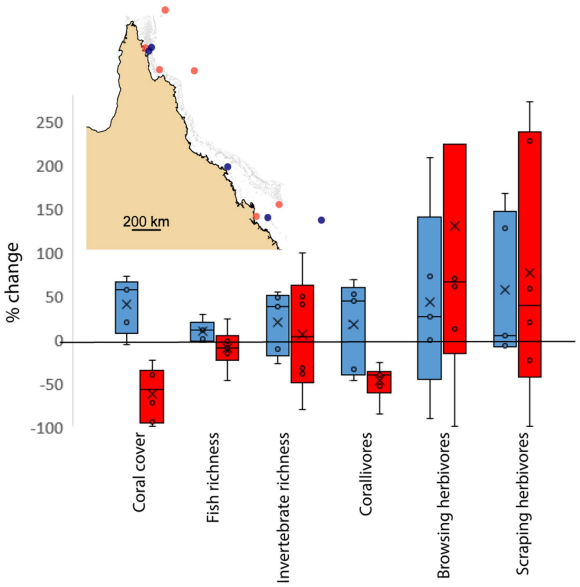
Extended Data Fig. 1 | Results of GLMMs for changes in coral and algal cover during the 2016 bleaching event. a, Changes in coral cover. **b**, Changes in algal cover. Change in cover is modelled as a function of the influences of starting cover (of corals and algae, respectively), wave exposure, thermal anomaly (DHD), the interaction between DHD and starting cover, depth category (depths between 4 and 10 m and >10 m modelled in comparison to <4-m depth), and the interaction between depth category and DHD ($n = 211$ site–depth combinations). All continuous predictors were normalized to mean = 0 and s.d. = 1 for comparative purposes.

© 2018 Springer Nature Limited. All rights reserved.



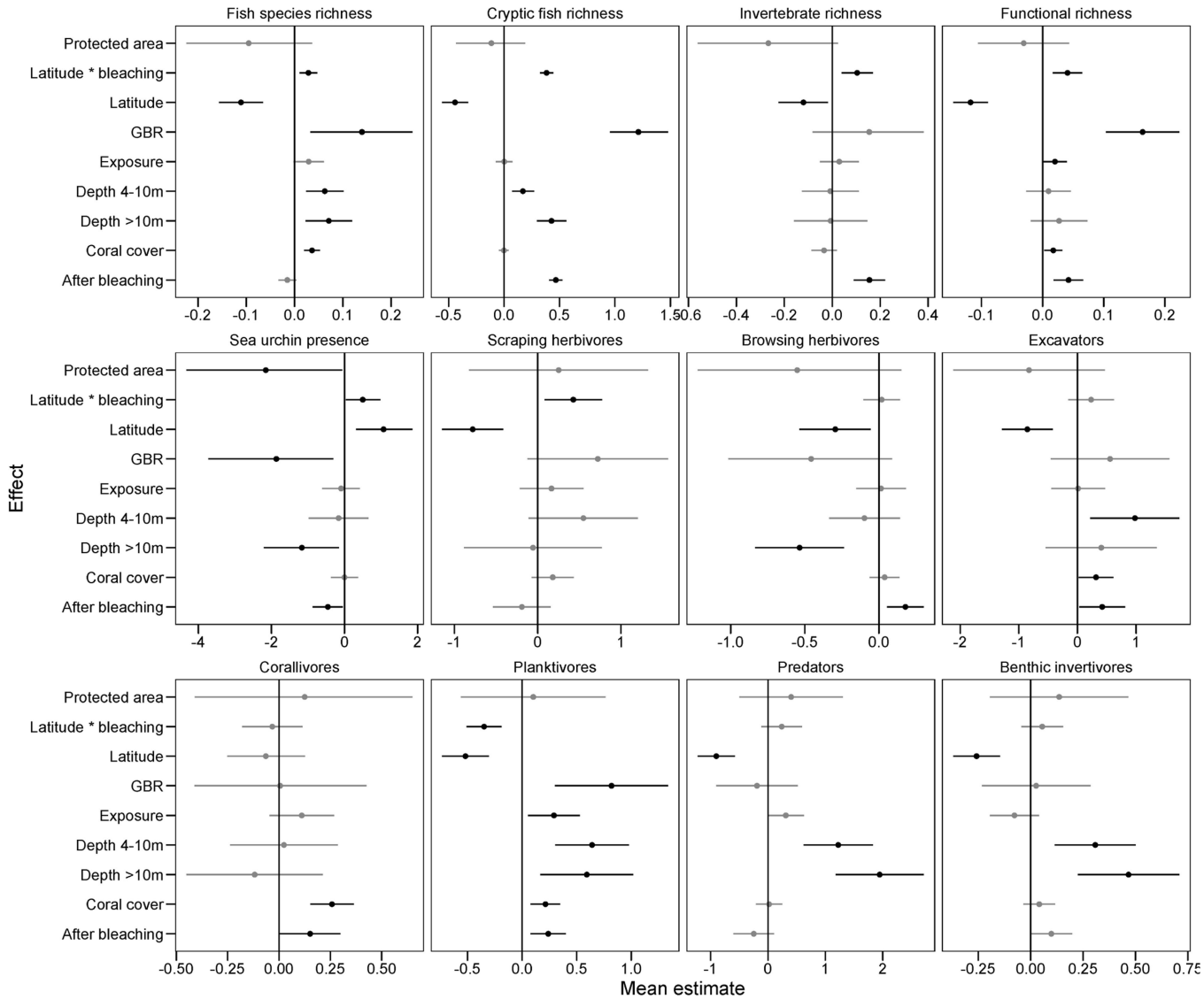
Extended Data Fig. 2 | Changes in algal and coral cover spanning the 2016 bleaching event. a, Coral- and algal-cover change were negatively correlated ($\rho = -0.56$). **b**, The greatest algal-cover increases occurred at sites with the lowest coral cover after the bleaching event ($\rho = -0.28$). $n = 211$ site–depth combinations.

© 2018 Springer Nature Limited. All rights reserved.



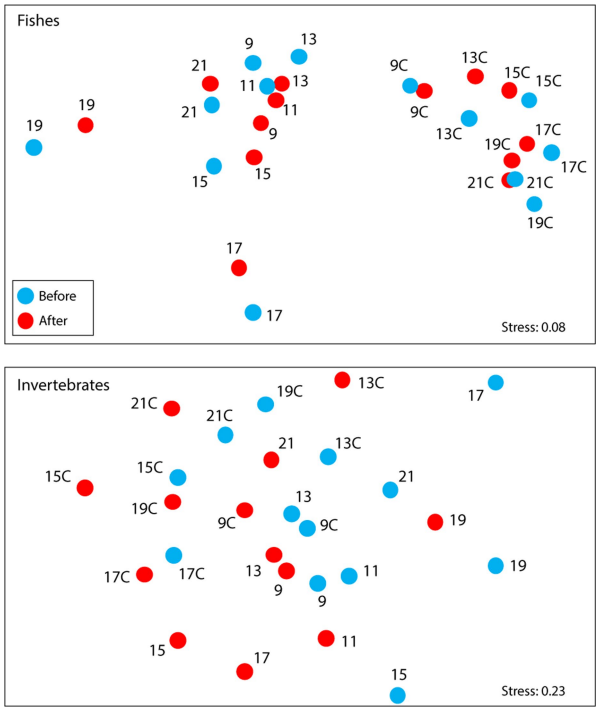
Extended Data Fig. 3 | Ecological changes on surveyed reefs most clearly affected by coral bleaching (red) versus un-impacted reefs (blue). Reefs categorized as bleached were those with >20% pre-bleaching live coral cover, that experienced >40 DHD and that lost >40% of pre-bleaching coral cover (see Methods for rationale). The un-impacted reefs were those that had >20% pre-bleaching live coral cover and experienced >40 DHD, but did not show a reduction in coral cover. The vertical axis is the percentage change of each metric across the reefs in each category ($n = 6$ bleached, $n = 5$ unbleached reefs), and horizontal lines on box plots show median, first and third quartiles, with the range indicated by the error bars. Crosses indicate means and circles indicate individual reefs within quartiles. Values for corallivores, browsing herbivores and scraping herbivores describe change in densities of species in these groups. Densities and species richness are means per 500 m^2 (fishes) or 100 m^2 (invertebrates). Bleached and unbleached reefs each include reefs from both northern and southern regions. Only coral cover differed noticably between these two groups of reefs (mean difference = −72%, with 95% credible intervals of 25–107%), although there was a small decline in corallivore densities post-bleaching (mean difference = 42% with 95% credible intervals from −0.24 to 78.0).

© 2018 Springer Nature Limited. All rights reserved.



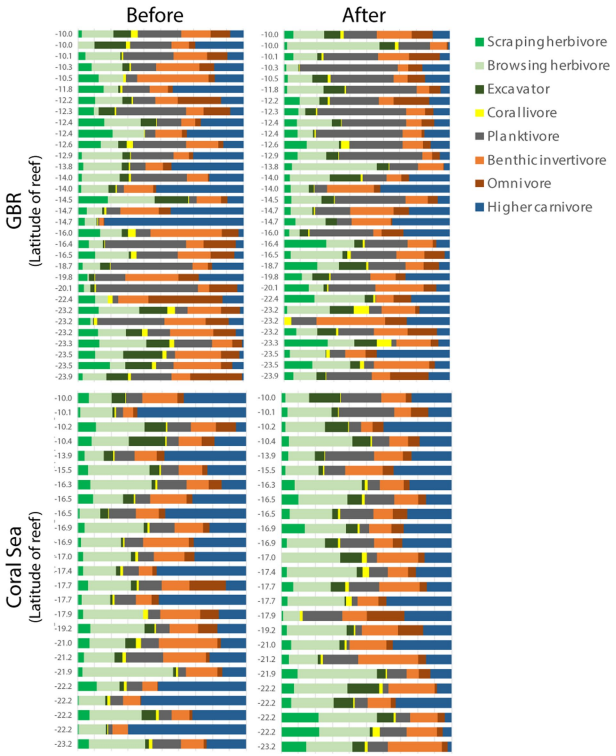
Extended Data Fig. 4 | Effect sizes from GLMMs of regional change for each ecological metric. Median additive effects of each covariate on the linear expectation for each metric (with 95% credible intervals as error bars) ($n = 233$ site-by-depth-category combinations). Effect sizes are on a log scale for all metrics, except for sea urchin presence, which gives the effect on the log-odds of presence versus absence. The influences of latitude, and its change from before to after the bleaching event (the interaction between latitude and bleaching (Latitude * bleaching)), are modelled in relation to differences between the GBR and Coral Sea reefs (GBR), wave exposure (Exposure), depth of the survey (depths between 4 and 10 m and >10 m modelled in comparison to <4-m depth), the percentage cover of live hard corals in the survey (Coral cover) and before versus after the bleaching event (After bleaching). Effects for which credible intervals do not overlap zero are indicated with black, rather than grey, points and error bars.

© 2018 Springer Nature Limited. All rights reserved.



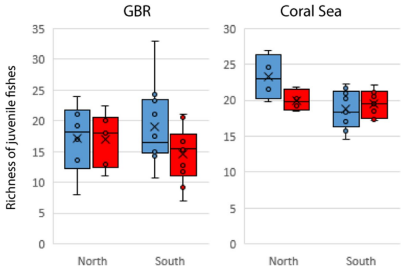
Extended Data Fig. 5 | Non-metric multidimensional scaling plots for reef fish and mobile invertebrate communities along the GBR and Coral Sea. Fish biomass data (top) and invertebrate abundance data (bottom) were averaged across surveys within 2° latitudinal bands, with number labels representing the northern latitude (that is, 21 represents the 2° band from 21° to 23° south). Coral Sea reefs are distinguished from those in the GBR by a 'C' in the label. Symbols have been colour-coded for data collected before and after the bleaching event ($n = 13$ latitudinal bands each before and after).

© 2018 Springer Nature Limited. All rights reserved.



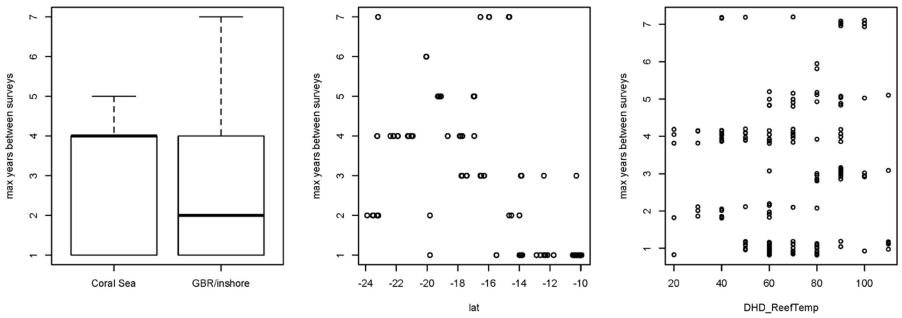
Extended Data Fig. 6 | Changes in the trophic structure of reef fishes following the 2016 mass bleaching event on the GBR and Coral Sea. Bars represent the proportion of total biomass made up by each trophic group, averaged across surveys on each reef, and reefs ordered by latitude. Cleaners and algal farmers were removed owing to their small contributions to biomass.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Local species richness of juvenile fishes (per 500 m^2) before and after the 2016 mass bleaching event on the GBR and Coral Sea. Species richness is shown before (blue) and after (red) the 2016 mass bleaching event on the GBR (left) and Coral Sea (right). 'North' reefs were north of 12° S ($n = 10$ reefs), and 'south' reefs were south of 19° S ($n = 19$ reefs). Juveniles were classified as any individuals 10 cm or less, for species that exceed 12.5 cm in maximum size. A Bayesian linear model indicated juvenile richness differed between the GBR and Coral Sea, but not between north and south or before and after the bleaching event (mean difference = 1.70 with lower and upper 95% credible intervals of −0.6 to 4.0). The distribution of raw data is shown in box plots, with crosses indicating means and circles indicate individual reefs within quartiles.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | The distribution of sampling effort through space and time. The temporal gap between pre- and post-bleaching surveys ($n = 768$ surveys total) between GBR and Coral Sea, along the latitudinal gradient, and locations experiencing different heating anomalies. For the box plot, the box shows the interquartile range and whiskers are 1.5× interquartile range.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Categories of coral and algal cover scored from photoquadrats

		Coral Sea		Great Barrier Reef	
Category	**Functional group**	**Before**	**After**	**Before**	**After**
Algae	*Caulerpa* spp.	3.6	4.6	4.7	11.4
	Crustose coralline algae	21.1	26.1	9.7	15.3
	Diatom/cyanobacterial slime	5.6	4.6	18.8	5.1
	Encrusting leathery algae	4.5	3.5	3.5	2.3
	Filamentous algae	2.2	1.6	3.7	1.4
	Foliose red algae	1.0	2.0	3.5	1.3
	Geniculate coralline algae	4.0	3.9	2.3	1.2
	Green calcified algae	16.7	20.0	2.5	3.9
	Other foliose green algae	4.9	8.5	1.0	17.0
	Small to medium foliose brown algae	1.5	15.4	4.6	4.0
	Turfing algae (<2 cm)	15.3	17.8	33.6	32.7
Coral	Ahermatypic corals	1.8	1.5	0	0
	Bleached coral	3.8	3.1	1.0	8.3
	Bottlebrush corals	5.7	16.2	4.5	3.2
	Branching *Acropora*	9.9	6.4	12.0	7.5
	Columnar corals	2.0	7.3	3.0	3.7
	Corymbose corals	2.9	2.8	4.3	5.1
	Digitate corals	3.8	1.4	1.9	2.2
	Encrusting corals	9.7	7.9	5.1	5.5
	Foliose corals	4.9	6.9	6.0	5.6
	Hydrocoral	2.3	5.4	3.9	1.3
	Large-polyp stony corals	1.5	1.2	1.8	1.3
	Massive corals	5.0	4.5	6.0	8.1
	Organ-pipe coral (*Tubipora*)	1.0	1.3	0	0
	Other branching/erect corals	5.1	8.7	7.3	10.9
	Pocillopora	3.3	3.2	2.6	3.1
	Soft corals and gorgonians	8.3	8.1	7.8	7.1
	Submassive corals	3.5	2.8	2.9	4.8
	Tabular Coral	4.7	3.4	9.6	12.8

Mean percentage cover values from before and after the bleaching event are shown for sites the Coral Sea and GBR. Categories are for live cover, with dead or bleached individuals or colonies scored as such (bleached corals were scored if white at the time of surveys, and only summed bleached corals were included here). Soft corals were excluded from summed cover for analysis of live hard coral.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0354-1

Differential tuning of excitation and inhibition shapes direction selectivity in ferret visual cortex

Daniel E. Wilson[1,2,3,4,5], Benjamin Scholl[1,5] & David Fitzpatrick[1]*

To encode specific sensory inputs, cortical neurons must generate selective responses for distinct stimulus features. In principle, a variety of factors can contribute to the response selectivity of a cortical neuron: the tuning and strength of excitatory[1–3] and inhibitory synaptic inputs[4–6], dendritic nonlinearities[7–9] and spike threshold[10,11]. Here we use a combination of techniques including in vivo whole-cell recording, synaptic- and cellular-resolution in vivo two-photon calcium imaging, and GABA (γ-aminobutyric acid) neuron-selective optogenetic manipulation to dissect the factors that contribute to the direction-selective responses of layer 2/3 neurons in ferret visual cortex (V1). Two-photon calcium imaging of dendritic spines[12,13] revealed that each neuron receives a mixture of excitatory synaptic inputs selective for the somatic preferred or null direction of motion. The relative number of preferred- and null-tuned excitatory inputs predicted a neuron's somatic direction preference, but failed to account for the degree of direction selectivity. By contrast, in vivo whole-cell patch-clamp recordings revealed a notable degree of direction selectivity in subthreshold responses that was significantly correlated with spiking direction selectivity. Subthreshold direction selectivity was predicted by the magnitude and variance of the response to the null direction of motion, and several lines of evidence, including conductance measurements, demonstrate that differential tuning of excitation and inhibition suppresses responses to the null direction of motion. Consistent with this idea, optogenetic inactivation of GABAergic neurons in layer 2/3 reduced direction selectivity by enhancing responses to the null direction. Furthermore, by optogenetically mapping connections of inhibitory neurons in layer 2/3 in vivo, we find that layer 2/3 inhibitory neurons make long-range, intercolumnar projections to excitatory neurons that prefer the opposite direction of motion. We conclude that intracortical inhibition exerts a major influence on the degree of direction selectivity in layer 2/3 of ferret V1 by suppressing responses to the null direction of motion.

Using sparse expression of the ultrasensitive fluorescent indicator protein GCaMP6s and two-photon calcium imaging[12,13], we first determined a neuron's somatic preferred direction using drifting grating stimuli (example cell, Fig. 1a) and then presented the somatic preferred and null directions while imaging spines on apical and basal dendrites. Individual dendritic branches contained spines that responded preferentially to either the somatic preferred or null directions (example branch, Fig. 1b); this diversity existed throughout the dendritic fields of single neurons (Fig. 1c). By quantifying spine responses using a direction selectivity index (DSI, see Methods) we found that spines tuned for the somatic preferred and null directions had different DSI values (Fig. 1d; Wilcoxon rank-sum, $P = 0.002$, $n = 384$ preferred spines, $n = 233$ null spines from $n = 17$ cells from 10 animals), and in 8 of 12 direction-selective cells (soma DSI > 0.3), more spines were tuned to the somatic preferred direction than to the null direction (Fig. 1e). Similar fractions of spines from basal (300 of 498) and apical (84 of 119) dendrites responded more strongly to the somatic preferred direction (60.2% and 70.6%, respectively). We then computed bootstrapped sums of normalized spine responses (see Methods) to assess the relation between the DSI of a neuron's inputs and of its soma. Summed excitatory synaptic input was weakly tuned (summed spine DSI $= 0.1 \pm 0.1$, median $\pm$ interquartile range (IQR), $n = 17$ cells) and we found no correlation between summed spine and somatic DSI across our sample (Fig. 1f; $r = -0.08$, $P = 0.75$, $n = 17$ cells), regardless of how we assessed synaptic input direction selectivity (see Methods, Extended Data Fig. 1a, b). Together, these data emphasize that the functional specificity of excitatory synaptic inputs converging onto individual layer 2/3 neurons is sufficient to account for somatic direction preference, but fails to explain the degree of somatic direction selectivity.

Previous work has suggested that spike threshold amplifies weak biases in excitatory synaptic inputs to enhance the spiking direction selectivity of neurons in layer 4 of cat V1[14]. To investigate whether subthreshold membrane potential (V_m) responses reflect the weak biases in excitatory inputs demonstrated by our spine imaging, we made whole-cell patch-clamp recordings from layer 2/3 neurons using a K^+-based internal solution ($n = 76$ cells from 23 animals; example, Fig. 2a, b). Most cells (78%, $n = 54$ of 69) showed strong spiking direction selectivity (Extended Data Fig. 2). Unexpectedly, nearly half of the cells with direction-selective spiking responses (48%, $n = 26$ of 54) showed strong direction tuning in their V_m (Fig. 2c), and the degree of V_m selectivity was correlated with spiking selectivity ($n = 69$ cells, $r = 0.56$, $P = 5.67 \times 10^{-7}$). The strong direction selectivity evident in V_m responses is in sharp contrast to the weak selectivity predicted by the distribution of excitatory synaptic inputs and forced us to consider factors that might contribute to the strong direction tuning of V_m responses. In principle, the emergence of strong subthreshold tuning from broadly tuned excitatory inputs could reflect mechanisms that enhance the effectiveness of excitatory inputs tuned to the preferred direction, diminish the effectiveness of excitatory inputs tuned to the null direction or a combination of these factors. To distinguish among these alternatives, we investigated whether there was a consistent relationship between subthreshold selectivity and subthreshold response amplitude to preferred and null direction stimuli. We found no correlation between V_m DSI and subthreshold response amplitude to the preferred direction (Fig. 2d; $r = -0.001$, $P = 0.99$, $n = 76$). By contrast, we found a strong anticorrelation between V_m DSI and null direction response amplitude (Fig. 2e; $r = -0.69$, $P = 5.07 \times 10^{-12}$, $n = 76$). These results indicate that factors that influence null direction responses are important in determining V_m selectivity.

We then considered the degree to which inhibitory inputs contribute to subthreshold responses to the null direction of motion[5,15,16]. Theoretical models predict that when levels of inhibition are high relative to excitation, not only is there a reduction in the level of depolarization, but also a reduction in the 'noise' or V_m variability[17,18]. We therefore examined the relationship between subthreshold DSI and V_m noise (see Methods) for the preferred direction and observed no significant correlation (Fig. 2f, $r = -0.10$, $P = 0.39$, $n = 76$ cells). Instead, we uncovered a strong anticorrelation between subthreshold DSI and V_m noise at the null direction (Extended Data Fig. 3; Fig. 2g, $r = -0.61$,

[1]Max Planck Florida Institute for Neuroscience, Jupiter, FL, USA. [2]IMPRS for Brain and Behavior, Florida Atlantic University, Jupiter, FL, USA. [3]Integrative Biology and Neuroscience Graduate Program, Florida Atlantic University, Jupiter, FL, USA. [4]Present address: Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [5]These authors contributed equally: Daniel E. Wilson, Benjamin Scholl. *e-mail: david.fitzpatrick@mpfi.org

© 2018 Springer Nature Limited. All rights reserved.

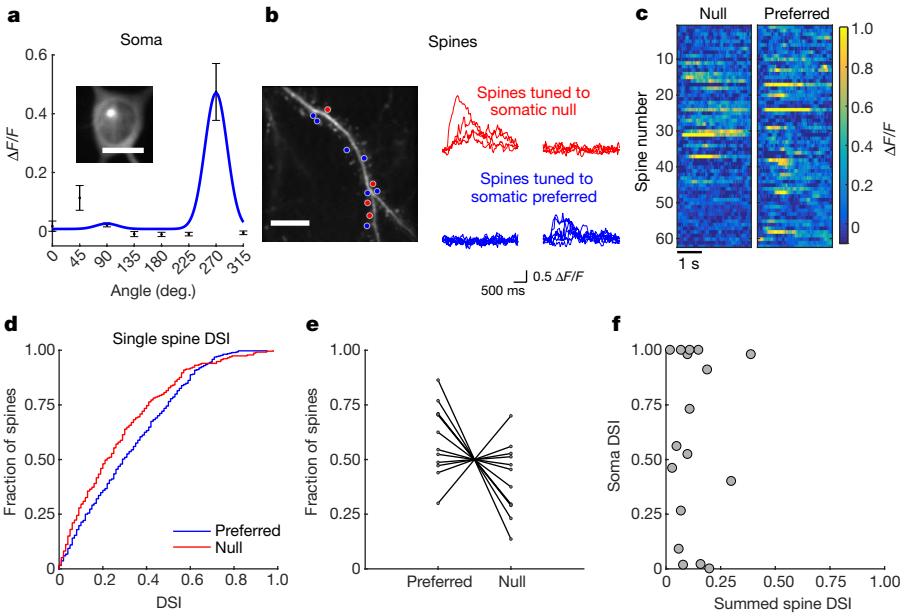


Fig. 1 | Direction tuning of excitatory synaptic inputs onto layer 2/3 neurons in ferret V1. a, Example soma tuning (data, mean ± s.e.m.; scale bar, 10 μm). **b**, Left, example dendritic spines ($n = 11$) pseudocoloured for direction preference (scale bar, 10 μm). Right, trial-averaged spine responses to somatic preferred or null directions. **c**, Trial-averaged responses for all significantly responsive spines ($n = 62$) from cell in **a**. **d**, DSI distributions for preferred ($n = 384$) and null ($n = 233$) direction-tuned spines. **e**, Fraction of spines on direction-selective cells ($n = 12$) preferring somatic preferred and null directions. **f**, Relationship between somatic DSI and summed spine DSI ($n = 17$ cells from 10 animals).

$P = 6.94 \times 10^{-9}$, $n = 76$ cells), consistent with a significant role for inhibition to the null direction in shaping the subthreshold direction selectivity of layer 2/3 neurons.

The idea that inhibition contributes to direction selectivity contrasts with a number of demonstrations that inhibition normalizes cortical activity[19,20]. These studies showed that excitation and inhibition are generally co-tuned[4,14,20,21], whereas our observations suggest that the tuning of excitatory and inhibitory inputs onto direction-selective neurons is dissimilar. To directly measure excitatory (G_e) and inhibitory (G_i) synaptic conductances underlying direction selectivity, we performed whole-cell patch-clamp recordings using a Cs^+-based internal solution[14] and recorded V_m responses to drifting gratings at different current steps to extract synaptic conductances and their direction tuning (Fig. 3a, b; see Methods). We observed a wide range of direction selectivity in synaptic conductances and found that excitatory and inhibitory DSIs were not correlated (Fig. 3c, $r = 0.043$, $P = 0.91$, $n = 10$ cells from 7 animals) and therefore not co-tuned. In half of our recorded neurons, excitation and inhibition preferred opposite directions ($\Delta\theta > 135°$) and across the population there was a significant bias towards preferring opposite directions (Extended Data Fig. 4a, Monte Carlo significance test, $P = 0.023$). Despite a lack of co-tuning, excitation and inhibition shared similar tuning bandwidth (Extended Data Fig. 4b).

Understanding the effect of inhibitory conductances on subthreshold responses requires consideration of co-occurring excitatory conductances. Thus, we measured the relative strength of inhibition as the ratio of inhibitory to excitatory conductance (G_i/G_e) and found that G_i/G_e was systematically larger for null direction than for preferred direction stimuli (Fig. 3e; Wilcoxon sign-rank, $P = 0.037$, $n = 10$). Moreover, the direction selectivity of predicted V_m from empirically measured synaptic conductances[22] (population tuning curves in Extended Data Fig. 4c) was significantly correlated with the G_i/G_e ratio at the null direction (Fig. 3f; $r = 0.81$, $P = 0.008$, $n = 10$ cells from 7 animals) but not the preferred direction (Extended Data Fig. 5). Consistent with our spine imaging data (Fig. 1f), predicted V_m direction selectivity was not correlated with the direction tuning of excitation alone (Fig. 3d; $r = 0.49$, $P = 0.15$, $n = 10$). Our measurements of synaptic conductances suggest that relatively stronger inhibitory input at the null direction enhances somatic direction selectivity.

If relatively greater G_i/G_e at the null direction contributes to direction selectivity in layer 2/3, inactivation of GABAergic neurons in layer 2/3 would reduce suppression at the null-direction and reduce selectivity, as suggested by previous pharmacological studies[23]. To test this hypothesis, we optogenetically suppressed layer 2/3 GABAergic neurons by expressing GtACR2 under the control of the mouse *Dlx5/6* (*mDlx*) enhancer and measured direction selectivity using whole-cell patch-clamp recordings with K^+-based internal solution. Optogenetic inactivation of GABAergic neurons (Extended Data Fig. 6) increased evoked response amplitude (Fig. 3g) and reduced V_m DSI (Fig. 3h; $P = 0.004$, Wilcoxon sign-rank, $n = 16$ cells from 4 animals) and spiking DSI (Extended Data Fig. 7a). Notably, changes in subthreshold direction selectivity were not related to the absolute V_m depolarization induced by GABAergic photoinhibition in individual neurons (Extended Data Fig. 7b). Instead, the degree to which null-direction responses were modulated by GABAergic suppression (see Methods) depended on the cell's V_m DSI (Fig. 3j; $r = 0.56$, $P = 0.025$, $n = 16$ cells), whereas no such relationship was observed for modulation of preferred-direction responses (Fig. 3i; $r = 0.20$, $P = 0.46$, $n = 16$ cells). On the basis of these results, we conclude that inhibition enhances subthreshold direction selectivity through null-direction suppression, and we would predict that GABAergic neurons preferring the opposite direction would contribute to this suppression.

GABAergic neurons in ferret V1 are direction-tuned and form direction columns[24] aligned with the underlying intrinsic signal direction preference map (Extended Data Fig. 8a, b). For GABAergic neurons to innervate oppositely tuned excitatory cells, their projections must extend beyond the local direction domain and into adjacent cortical columns. This would be inconsistent with studies from mouse V1, where excitatory neurons receive inhibitory input from local (within 100–200 μm) GABAergic neurons[25–27]. However, in carnivore V1, it has been shown that GABAergic neurons make axonal projections that span longer distances[28,29]. To test whether GABAergic neurons project beyond their local cortical columns, we labelled axon projections with punctate injections of AAV2/1-mDlx-GCaMP6s and characterized the direction tuning of axon projections at sites distal to the injection location (Fig. 4a–d). A substantial fraction (60.5%) of long-range projecting individual boutons exhibited direction-selective responses

© 2018 Springer Nature Limited. All rights reserved.

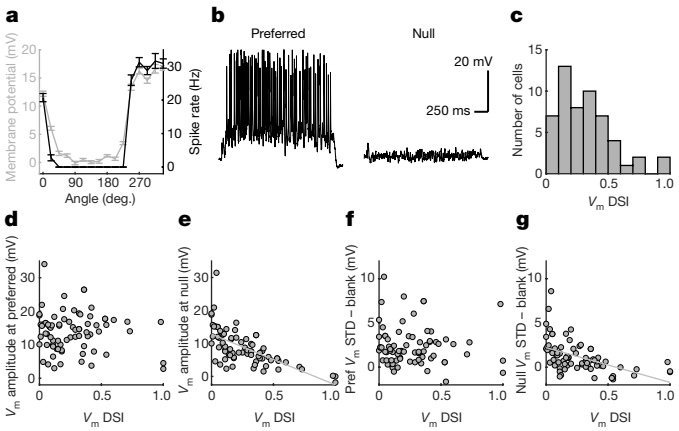


Fig. 2 | Subthreshold direction selectivity and evidence for null direction suppression. a, Example direction tuning of subthreshold V_m (grey) and spiking responses (black). Data are mean ± s.e.m. **b**, Example single trial responses. **c**, Distribution of V_m DSI for 54 cells with direction-selective spiking (DSI > 0.3). **d**, No relationship between preferred direction response and V_m DSI ($n = 76$ from 23 animals). **e**, Relationship between null direction response and V_m DSI; grey line is least-squares fit ($n = 76$). **f**, No relationship between preferred direction V_m noise (defined as the difference in V_m s.d. between a given stimulus and the blank) and V_m DSI ($n = 76$). **g**, Relationship between null direction V_m noise and V_m DSI; grey line is least-squares fit ($n = 76$).

(Fig. 4e; population median 0.39 ± 0.46 IQR, $n = 815$ boutons, 8 fields of view from 2 animals) and diverse direction preferences when compared to the intrinsic signal map (Fig. 4d). We found an unexpected abundance of direction-selective GABAergic boutons tuned to the direction opposite to direction domains (Fig. 4f). Furthermore, individual bouton preferences were significantly different from the map (Monte Carlo significance test, $P < 0.001$, $n = 493$ boutons), providing an anatomical substrate for the synaptic inhibition observed in our previous measurements.

To examine whether individual neurons receive inhibitory synaptic input from distant GABAergic neurons, we developed a technique called somatically targeted optogenetic membrane potential mapping (STOMPM) to directly map the spatial connectivity of inhibitory neurons onto excitatory neurons in vivo. We localized channelrhodopsin-2 to the soma and proximal dendrites of GABAergic neurons using a Kv2.1 targeting motif (Fig. 4g) to prevent stimulation of the neuropil and to enhance our functional resolution. As the direction preferences of GABAergic neurons are smoothly mapped in a columnar fashion[24] (Extended Data Fig. 8), we used patterned photostimulation driven by a digital light processing projector[30] to activate GABAergic neurons in local cortical regions (~100–200 μm, Fig. 4g) while recording from single neurons to measure inhibitory postsynaptic potentials (IPSPs) (example cell, Extended Data Fig. 9). Optical stimulation evoked robust IPSPs (Fig. 4h) even at spots distant from recorded cells. Neurons received inhibitory synaptic input from long distances (Fig. 4i); inhibitory input fields often exceeded 1 mm along their major axes (Fig. 4i–k; major axis length 930 ± 278 μm, median ± IQR, $n = 21$ cells from 7 animals) and many inputs arrived from distances greater than 500 μm (Extended Data Fig. 10). We recognize that these measures are likely to underestimate the total extent of input field size (see Methods). Finally, we aligned our stimulation grid with the intrinsic signal direction preference map (Fig. 4l) to characterize the functional origin of evoked IPSPs. Neurons with direction-selective V_m ($n = 7$ cells, mean tuning curve in Fig. 4m) received almost equivalent inhibitory synaptic input from null-tuned as from preferred-tuned direction domains (Fig. 4n).

Previous studies suggest that inhibition and excitation are generally co-tuned[4,14,20,21,31] (Fig. 4o), as shown for orientation selectivity in mouse V1[4] and layer 4 simple cells of cat V1[11], albeit with distinct temporal dynamics[4], acting to scale or gate overall responses[22].

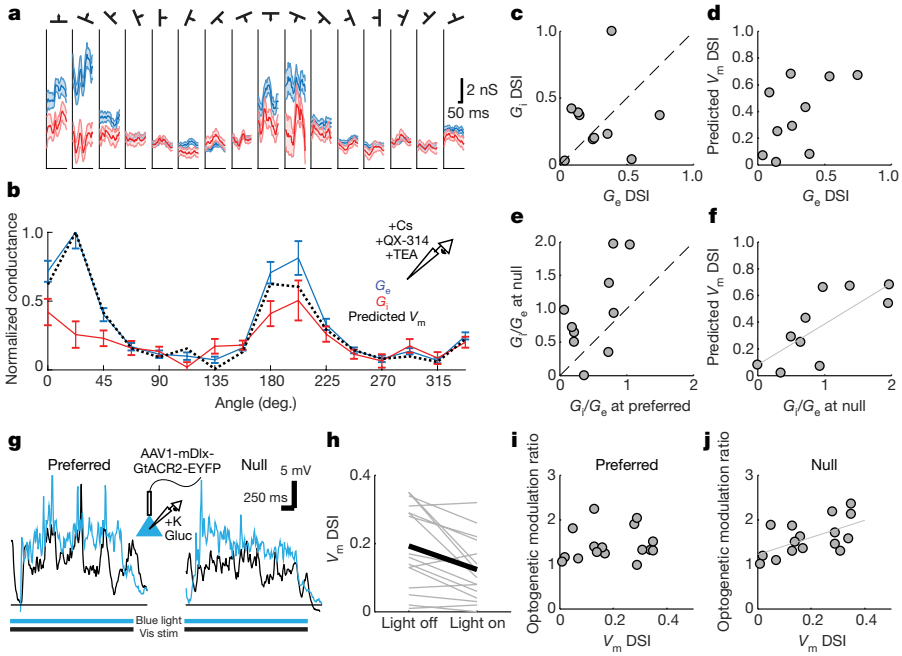


Fig. 3 | Differential tuning between excitation and inhibition enhances direction selectivity. a, Estimated excitatory (blue, G_e) and inhibitory (red, G_i) synaptic conductances driven by gratings from an example cell; line is bootstrapped mean and error bars are bootstrapped s.d. **b**, Tuning of peak (see Methods) synaptic conductances and predicted V_m (dashed line) for cell in **a**; data are bootstrapped mean and s.d. **c**, Comparison of G_e and G_i DSI ($n = 10$ from 7 animals). **d**, Predicted V_m DSI (see Methods) compared to G_e DSI ($n = 10$). **e**, Comparison of G_i/G_e at null and preferred directions ($n = 10$). **f**, Predicted V_m DSI compared to null direction G_i/G_e; grey line is least-squares fit ($n = 10$). **g**, Example V_m during visual stimulation and inactivation of GABAergic neurons expressing GtACR2 (cyan) or without inactivation (black); dashed line is resting V_m. **h**, Comparison of V_m DSI with and without inactivation; black line indicates population means ($n = 16$ from 4 animals). **i**, Optogenetic modulation of preferred direction response versus V_m DSI ($n = 16$). **j**, Optogenetic modulation of null direction response versus V_m DSI; grey line is least-squares fit ($n = 16$).

© 2018 Springer Nature Limited. All rights reserved.

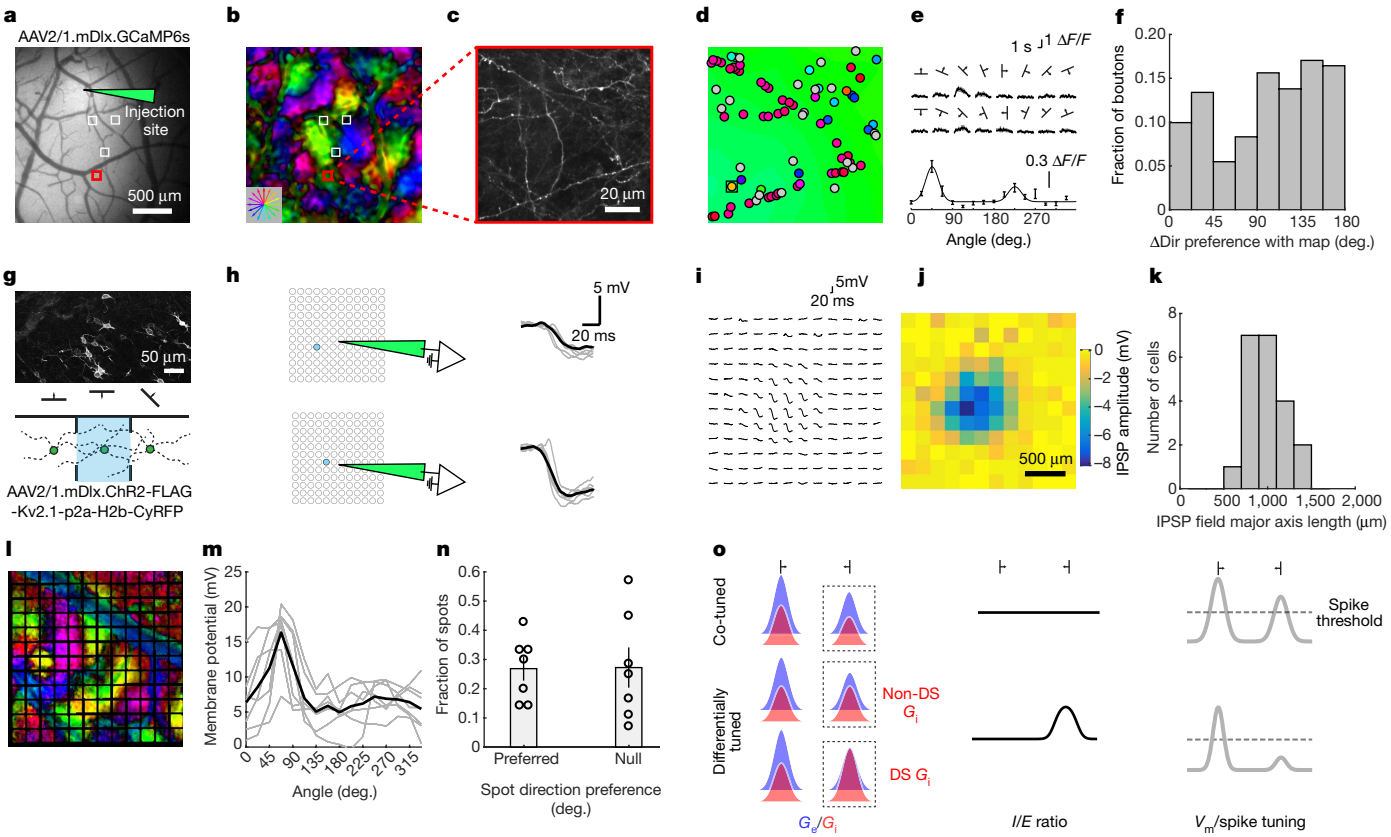


Fig. 4 | Inhibitory interneurons make long-range, intercolumnar projections onto excitatory neurons. a, Epifluorescence image of injection site with GABAergic axon imaging sites highlighted; example field of view (FOV) in red. **b**, Intrinsic signal polar direction map for **a**. **c**, Example bouton FOV. **d**, GABAergic boutons overlaid on direction preference map; direction preference of boutons and intrinsic signal pseudocoloured as in **b**, bi-directional boutons coloured grey. **e**, Example bouton tuning curve (box in **d**); data are mean ± s.e.m. **f**, Distribution of direction preference difference between GABAergic boutons ($n = 493$) and corresponding intrinsic signal direction preference map. **g**, Top, Flag staining of cells expressing AAV1-mDlx-ChR2-Flag-Kv2.1-p2a-H2b-CyRFP; bottom, experimental design: neurons in different cortical columns are optogenetically activated. **h**, Example of single spot illumination and V_m responses. **i**, Mean IPSP waveforms evoked by sampled spots. **j**, Map of IPSP amplitudes. **k**, Distribution of IPSP-field major axis lengths across cells ($n = 21$). **l**, Example stimulation grid aligned to intrinsic signal polar direction map. **m**, Peak-aligned average direction tuning curve for cells with direction-tuned membrane potential (DSI > 0.3, black, individual cells in grey, $n = 7$). **n**, Fraction of spots tuned to a cell's preferred (<45°) or null (>135°) direction (grey bars, mean ± s.e.m.), **o**, Cartoon model of co-tuning (top) and differential tuning (bottom) of excitation (G_e) and inhibition (G_i) for direction. Subthreshold direction selectivity is inherited from synaptic conductances when co-tuned. Differential tuning of G_e and G_i, whereby there is greater G_i/G_e at the null direction, can preferentially suppress excitation and enhance subthreshold selectivity. With differential tuning, inhibition can be either bidirectional or oppositely tuned for direction relative to G_e.

By contrast, we find that cortical inhibition can suppress responses to specific stimuli through differential tuning with excitation (Fig. 4o). Such differential tuning can arise through multiple combinations of excitation and inhibition, such that null-direction suppression is driven by either null-biased or equivalent inhibitory inputs for both directions (Fig. 4o). Differential tuning can enhance subthreshold selectivity, which is further augmented through spike threshold[11] (Fig. 4o). Our findings are conceptually similar to those in retinal ganglion cells[32], but differ in exact circuit implementation as ganglion cell direction selectivity arises through inhibitory input mediated by starburst amacrine cells. One factor we did not consider is the temporal interplay between excitation and inhibition, which could have an important contributing role in enhancing selectivity[31]. Together with the results of previous studies, our findings indicate that the selective responses of cortical neurons are built with a broadly tuned palette of excitatory synaptic inputs that is further refined by enhancing responses to the preferred stimulus[7,8,12] and suppressing responses to non-preferred stimuli.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0354-1

Received: 6 December 2017; Accepted: 8 June 2018;
Published online 25 July 2018.

1. Cossell, L. et al. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* **518**, 399–403 (2015).
2. Ko, H. et al. Functional specificity of local synaptic connections in neocortical networks. *Nature* **473**, 87–91 (2011).
3. Lee, W. C. et al. Anatomy and function of an excitatory network in the visual cortex. *Nature* **532**, 370–374 (2016).
4. Tan, A. Y., Brown, B. D., Scholl, B., Mohanty, D. & Priebe, N. J. Orientation selectivity of synaptic input to neurons in mouse and cat primary visual cortex. *J. Neurosci.* **31**, 12339–12350 (2011).
5. Monier, C., Chavane, F., Baudot, P., Graham, L. J. & Frégnac, Y. Orientation and direction selectivity of synaptic inputs in visual cortical neurons: a diversity of combinations produces spike tuning. *Neuron* **37**, 663–680 (2003).
6. Liu, B. H. et al. Broad inhibition sharpens orientation selectivity by expanding input dynamic range in mouse simple cells. *Neuron* **71**, 542–554 (2011).
7. Smith, S. L., Smith, I. T., Branco, T. & Häusser, M. Dendritic spikes enhance stimulus selectivity in cortical neurons *in vivo*. *Nature* **503**, 115–120 (2013).
8. Palmer, L. M. et al. NMDA spikes enhance action potential generation during sensory input. *Nat. Neurosci.* **17**, 383–390 (2014).
9. Lavzin, M., Rapoport, S., Polsky, A., Garion, L. & Schiller, J. Nonlinear dendritic processing determines angular tuning of barrel cortex neurons *in vivo*. *Nature* **490**, 397–401 (2012).
10. Priebe, N. J. & Ferster, D. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron* **57**, 482–497 (2008).
11. Anderson, J. S., Carandini, M. & Ferster, D. Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *J. Neurophysiol.* **84**, 909–926 (2000).

© 2018 Springer Nature Limited. All rights reserved.

12. Wilson, D. E., Whitney, D. E., Scholl, B. & Fitzpatrick, D. Orientation selectivity and the functional clustering of synaptic inputs in primary visual cortex. *Nat. Neurosci.* **19**, 1003–1009 (2016).
13. Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
14. Priebe, N. J. & Ferster, D. Direction selectivity of excitation and inhibition in simple cells of the cat primary visual cortex. *Neuron* **45**, 133–145 (2005).
15. Torre, V. & Poggio, T. A synaptic mechanism possibly underlying directional selectivity to motion. *Proc. R. Soc. Lond. B* **202**, 409–416 (1978).
16. Sillito, A. M. Inhibitory processes underlying the directional specificity of simple, complex and hypercomplex cells in the cat's visual cortex. *J. Physiol. (Lond.)* **271**, 699–720 (1977).
17. Prescott, S. A. & De Koninck, Y. Gain control of firing rate by shunting inhibition: roles of synaptic noise and dendritic saturation. *Proc. Natl Acad. Sci. USA* **100**, 2076–2081 (2003).
18. Chance, F. S., Abbott, L. F. & Reyes, A. D. Gain modulation from background synaptic input. *Neuron* **35**, 773–782 (2002).
19. Katzner, S., Busse, L. & Carandini, M. GABAA inhibition controls response gain in visual cortex. *J. Neurosci.* **31**, 5931–5941 (2011).
20. Isaacson, J. S. & Scanziani, M. How inhibition shapes cortical activity. *Neuron* **72**, 231–243 (2011).
21. Wehr, M. & Zador, A. M. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426**, 442–446 (2003).
22. Atallah, B. V., Bruns, W., Carandini, M. & Scanziani, M. Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. *Neuron* **73**, 159–170 (2012).
23. Crook, J. M., Kisvárday, Z. F. & Eysel, U. T. Evidence for a contribution of lateral inhibition to orientation tuning and direction selectivity in cat visual cortex: reversible inactivation of functionally characterized sites combined with neuroanatomical tracing techniques. *Eur. J. Neurosci.* **10**, 2056–2075 (1998).
24. Wilson, D. E. et al. GABAergic neurons in ferret visual cortex participate in functionally specific networks. *Neuron* **93**, 1058–1065.e4 (2017).
25. Packer, A. M. & Yuste, R. Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: a canonical microcircuit for inhibition? *J. Neurosci.* **31**, 13260–13271 (2011).
26. Hofer, S. B. et al. Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex. *Nat. Neurosci.* **14**, 1045–1052 (2011).
27. Levy, R. B. & Reyes, A. D. Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. *J. Neurosci.* **32**, 5609–5619 (2012).
28. Martin, K. A., Somogyi, P. & Whitteridge, D. Physiological and morphological properties of identified basket cells in the cat's visual cortex. *Exp. Brain Res.* **50**, 193–200 (1983).
29. Roerig, B. & Chen, B. Relationships of local inhibitory and excitatory circuits to orientation preference maps in ferret visual cortex. *Cereb. Cortex* **12**, 187–198 (2002).
30. Dhawale, A. K., Hagiwara, A., Bhalla, U. S., Murthy, V. N. & Albeanu, D. F. Non-redundant odor coding by sister mitral cells revealed by light addressable glomeruli in the mouse. *Nat. Neurosci.* **13**, 1404–1412 (2010).
31. Li, Y. T., Liu, B. H., Chou, X. L., Zhang, L. I. & Tao, H. W. Strengthening of direction selectivity by broadly tuned and spatiotemporally slightly offset inhibition in mouse visual cortex. *Cereb. Cortex* **25**, 2466–2477 (2015).
32. Borst, A. & Helmstaedter, M. Common circuit design in fly and mammalian motion vision. *Nat. Neurosci.* **18**, 1067–1076 (2015).

Acknowledgements We thank D. Whitney for help with analysis, A. Banerjee and F. Albeanu for help with patterned photostimulation setup, J. Chang for discussions about optogenetic stimulation, D. Ouimet for surgical assistance, N. Shultz and R. Satterfield for histology, T. Laviv for discussions about CyRFP, C. Baker and M. Bolton for cloning and the gift of the ChR2 construct, and the GENIE project for access to GCaMP6. This work was supported by EY011488, the Max Planck Florida Institute for Neuroscience, and the Max Planck Society.

Reviewer information *Nature* thanks A. Huberman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.E.W., B.S. and D.F. conceived experiments. D.E.W. and B.S. performed experiments and analysed data with guidance from D.F. D.E.W., B.S. and D.F. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0354-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0354-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.F.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

All procedures were performed according to NIH guidelines and were approved by the Institutional Animal Care and Use Committee at Max Planck Florida Institute for Neuroscience.

Constructs. pAAV-mDlx-GCaMP6f-Fishell-2[33] was a gift from G. Fishell (Addgene plasmid # 83899). pFUGW-hGtACR2-EYFP[34] was a gift from J. Spudich (Addgene plasmid # 67877). pCAG-CyRFP1 was a gift from R. Yasuda (Addgene plasmid # 84356). pAAV.mDlx.ChR2-Flag-Kv2.1.p2a.H2b-CyRFP was a gift from M. Bolton.

Virus injection. Female ferrets aged about 30 days (~P30; Marshall Farms) were anaesthetized with ketamine (50 mg/kg) and isoflurane (1–3%, delivered in O_2), then intubated and artificially respirated. Atropine was administered to reduce secretions and a 1:1 mixture of lidocaine and bupivacaine administered subcutaneously into the scalp. Animals were placed on a feedback-controlled heating pad to maintain internal temperature at 37 °C. Under sterile surgical conditions, a small craniotomy (0.8 mm diameter) was made over the visual cortex 7–8 mm lateral and 2–3 mm anterior to the lambda.

For spine imaging, we injected (52 nl per depth) a mixture (1:100,000) of AAV1-hSyn-Cre and AAV1-Syn-FLEX-GCaMP6s (UPenn, ~1×10^{13} GC/ml) at 400 and 200 μm below the pia through bevelled glass micropipettes (10–15 μm outer diameter). For imaging GABAergic axons or somata, we injected 5-30 nl of AAV1-mDlx-GCaMP6s at 400 and 200 μm below the pia. For optogenetic inactivation experiments, we injected 1 μl of AAV1-mDlx-GtACR2-eYFP (titre $>1 \times 10^{13}$ GC/ml, custom preparation from Vigene) at 400 and 200 μm below the pia through beveled glass micropipettes (15-20 μm outer diameter). For STOMPM (see below), we injected 1 μl of AAV1-mDlx-ChR2-Flag-Kv2.1-p2a-H2b-CyRFP (titre $>1 \times 10^{13}$ GC/ml, custom preparation from Vigene) through bevelled glass micropipettes (15–20 μm outer diameter). To prevent dural regrowth and adhesion, the craniotomy was filled with sterile 1% w/v agarose (Type IIIa, Sigma-Aldrich).

Cranial window. After 3–5 weeks of expression, ferrets were anaesthetized with 50mg/kg ketamine and 1–3% isoflurane. Atropine and bupivacaine were administered as in virus injection procedure. Animals were placed on a feedback-controlled heating pad to maintain an internal temperature of 37–38 °C. A tracheotomy was performed and an endotracheal tube installed to artificially respirate the animal. Isoflurane was delivered between 1 and 3% throughout the surgical procedure to maintain a surgical plane of anaesthesia. An intravenous cannula was placed to deliver fluids. ECG, end tidal CO2, external temperature and internal temperature were continuously monitored throughout the imaging session.

The scalp was retracted and a custom titanium headplate adhered to the skull using C&B Metabond (Parkell). A 6–7-mm craniotomy was performed at the viral injection site and the dura retracted to reveal the cortex. For spine and axon imaging, one to two pieces of custom coverglass (3 mm diameter, 0.7 mm thickness, Warner Instruments) were adhered to a larger coverglass (8 mm diameter, #1.5 thickness, Electron Microscopy Sciences) using optical adhesive (# 71, Norland Products) and placed onto the brain to gently compress the underlying cortex and dampen biological motion during imaging. For population imaging, a single coverglass (5 mm diameter, #1.5 thickness, Electron Microscopy Sciences) was adhered to the bottom of a titanium insert and then placed onto the brain. In both cases, a stainless steel retaining ring (5/16-inch internal retaining ring, McMaster-Carr) maintained downward pressure on the cranial window throughout the experiment.

For whole-cell recording and optogenetic experiments, the cranial window was filled with agarose (1.6%w/v, Type IIIa, Sigma) and a coverglass placed on top of the agarose. For pipette access, we drilled holes offset from the centre of the coverglass to allow our pipette to approach the centre of the cranial window at an oblique angle. The cranial window was hermetically sealed using a stainless steel retaining ring (5/16-inch internal retaining ring, McMaster-Carr), Kwik-Cast (World Precision Instruments), and Vetbond (3M). A 1:1 mixture of 1% tropicamide ophthalmic solution (Akorn) and 10% phenylephrine hydrochloride ophthalmic solution (Akorn) was applied to both eyes to dilate the pupils and retract the nictating membranes. Contact lenses were inserted to protect the eyes. Upon completion of the surgical procedure, isoflurane was gradually reduced (0.6 to 1.5%) and then vecuronium (2 mg kg^{-1} hr^{-1}) or pancuronium (2 mg kg^{-1} hr^{-1}) was delivered intravenously to immobilize the animal.

Visual stimuli. Visual stimuli were generated using Psychopy[35]. The monitor was typically placed 25 cm from the animal. After mapping somatic direction tuning using a grating protocol, we presented the somatic preferred and null directions of motion while imaging dendrites and dendritic spines. For whole-cell recording, we optimized the preferred spatial frequency of the stimulus for the cell being recorded. Typical preferred spatial frequencies ranged from 0.04 to 0.25 cycles per degree.

Two-photon imaging. Two-photon imaging was performed on a Bergamo II microscope (Thorlabs) running Scanimage[36] 2015 or 2016 (Vidrio Technologies) with 940-nm dispersion-compensated excitation provided by an Insight DS+ (Spectraphysics). For spine and axon imaging, power after the objective was limited to <60 mW, dependent on depth. Cells were selected for imaging on the basis of their position relative to large blood vessels, responsiveness to visual stimulation, and lack of prolonged calcium transients resulting from overexpression of GCaMP6s. Images were collected at 30 Hz using bidirectional scanning with 512×512 pixel resolution. Images of somata ranged from 50 to 100 μm on a side, while images of dendrites were ~40 μm on a side. Images of axons were collected at 512×512 pixel resolution with fields of view ~100 μm on a side.

Whole-cell patch-clamp recordings. Recordings were performed by inserting a pipette through an agarose-filled craniotomy or by using a coverglass with a hole drilled for pipette access. A silver–silver chloride reference electrode was inserted into the agarose or muscle. Recordings were made in current clamp mode and current pulses delivered by custom Labview software.

For measurements of membrane potential tuning, spike tuning, effects of optogenetic inhibition and connectivity mapping, pipettes of 5–8 MΩ resistance were pulled using borosilicate glass (King Precision Glass) and filled with an intracellular solution containing (in mM) 135 K gluconate, 4 KCl, 10 HEPES, 10 Na_2-phosphocreatine, 4 Mg-ATP, 0.3 Na_3GTP, 0–0.1 Alexa 594 or 488, pH 7.2, 295 mOsm. Neurons were recorded from layer 2/3 (100 to 800 μm below the pia) using a Multiclamp 700B (Molecular Devices). Series resistance and pipette capacitance were corrected online. Series resistance for recordings typically ranged from 20 to 80 MΩ. Analogue signals were digitized using Spike2 (CED). For optogenetic inactivation experiments, a fibre (1 mm, NA .63) coupled to a 455 nm LED light source (Prizmatix) was lowered to 3–5 mm above the cortical surface. Power density at the cortical surface ranged from 1 to 4 mW/mm^2. Optogenetic stimulation either coincided with visual stimulation, or began with a brief ramp (100 to 300 ms) before visual stimulation.

For measurements of synaptic conductances, the internal solution contained (in mM) 135 Cs-$MeSO_4$, 10 QX-314, 4 TEA-Cl, 2 EGTA, 2 MgATP, 10 HEPES, 10 Na_2-phosphocreatine (pH 7.3, 295 mOsm) and pipettes were typically 6–9 MΩ. Capacitance compensation was corrected online and series resistance was corrected online or offline. Conductance measurements typically began around 30 min after break-in to allow the internal solution of the pipette to dialyse the cell, eliminating action potentials and depolarizing the resting membrane potential as expected with the use of Cs^+ and QX-314.

Connectivity mapping. Connectivity mapping (STOMPM) was performed on a custom-built microscope based on previously published designs[30]. A digital light processing projector (X600, Optoma) with its colour wheel removed was mounted to a tilt platform (Siskiyou) and linear stage (Thorlabs). A 50 mm f/1.4 SLR lens (Nikkor) was mounted as close as possible to the projector and coupled to an achromatic doublet (AC508-150-A). Light passed through a blue dichroic filter (52-532, Edmund Optics) and was reflected onto the sample using a dichroic mirror (T495LPXR, Chroma), and focused onto the sample using a 35 mm f/2.0 SLR lens (Nikkor). Emission light passed through a 105 mm f2.0 lens (Nikkor) and an emission filter (FF01-600) and was imaged onto a camera (Xyla, Andor) controlled by Micromanager (http://www.micro-manager.org). Single pixels on the DMD corresponded to ~4 μm at the sample. Diffuse background light was <0.1 mW/mm^2. Opsin was restricted to the soma using the Kv2.1 targeting motif[37,38]. Before obtaining whole-cell recordings, we focused excitation light on the cortical surface. Upon break-in, we first measured the direction tuning of the cell using a grating protocol. Then, we centred a stimulation grid on the pipette and delivered 25–50 trials of random grid stimulation. Spots were typically 100–200 μm full width at half-maximum (FWHM), 1–3 mW power, and displayed for 100 ms. We used positive current injection to depolarize the cell and increase the driving force for IPSPs (reversal potential ~75–70 mV, Extended Data Fig. 9). We probably underestimate input field sizes owing to limitations in the spatial spread of virus injection, blue light absorption in blood vessels and experimental geometry in which the large patch pipette interferes with light stimulation.

Intrinsic signal imaging. Intrinsic signal imaging was performed on the STOMPM microscope or on the Thorlabs Bergamo II. The cortex was illuminated with blue light to obtain a blood vessel map, after which collimated 630 nm light from an LED (Thorlabs) was directed onto the surface of the brain to measure intrinsic haemodynamic responses. Visually evoked responses were collected at ~50 Hz using an Andor Xyla camera. Visual stimuli were blockwise grating stimuli (8 s on, 8 s off, 0.06–0.1 cycles per degree, 16 directions).

Fixation and immunostaining. Upon completion of imaging, isoflurane was raised to 5% and 0.5 ml Euthasol given IV. The animal was transcardially perfused with 100 ml of 0.9% NaCl (w/v) and then 500 ml of 4% paraformaldehyde in 0.1 M phosphate buffer (PB). The brain was dissected and post-fixed overnight in 4% PFA in 0.1M PB at 4 °C. Cryoprotection was carried out in 30% sucrose for 2 days, at which time tissue was sliced at 50 um on a Leica SM 2010R. Cryosections were rinsed in 0.1M phosphate buffered saline (PBS), blocked in blocking solution containing 1% BSA, 2% normal goat serum, and 0.3% triton X-100 in PBS for 1 h, and then incubated in mouse anti-FlagM2 at 1:500 (Sigma cat# F1804) overnight at room temperature. After three washes in buffer, 10 min each, the sections were

© 2018 Springer Nature Limited. All rights reserved.

incubated in Alexa goat anti-mouse 488 at 1:500 (Thermo Fisher, cat# A32723) for 2 h at room temperature. After incubation, sections were washed twice in 0.1M PBS for 10 min each, followed by one wash in 0.1M PB. Sections were then mounted on Superfrost Plus slides (VWR, West Chester, PA) and coverslipped with Slowfade Gold (Thermo Fisher cat# S36936). Control slides were treated without the primary antibody. These control sections showed no labelling. To test the specificity of secondary antibodies, the secondary antibody was applied to the tissue without a primary antibody; no staining was observed in these controls.

Analysis. *Calcium imaging.* Imaging data were excluded from analysis if motion along the *z*-axis was detected. Dendrite images were corrected for in-plane motion via a 2D cross-correlation based approach in MATLAB. Axon images were corrected for in-plane motion using a piecewise non-rigid motion correction algorithm[39]. ROIs were drawn in ImageJ; dendritic ROIs spanned contiguous dendritic segments and bouton/spine ROIs were circular. Mean pixel values for single ROIs were computed over the imaging time series and imported into MATLAB via MIJ[40]. $\Delta F/F_0$ was computed by defining F_0 using a 60 s percentile filter (typically 10th percentile), which was then low-pass filtered at 0.01 Hz. Bouton and somatic responses were computed as the average response to the visual stimulus and were included for analysis of direction selectivity if $\Delta F/F_0$ exceeded 10% and 1–CV was >0.1. $\Delta F/F_0$ traces were median filtered with a 3-sample window. For spine signals, we subtracted a scaled version of the dendritic signal to remove backpropagating action potentials as previously performed[12,13]. $\Delta F/F_0$ traces were synchronized to stimulus triggers sent from Psychopy and collected by Spike2. Spines were included for analysis if the average response exceeded 2 median absolute deviations above the baseline noise (measured during the blank) and were weakly correlated with the dendritic signal (Spearman's correlation, $r < 0.3$). Some spine traces contained negative events after subtraction, so we excluded negative $\Delta F/F$ values when computing Spearman's correlation between the spine and the dendrite. Because the amplitude of NMDA receptor mediated calcium transients are not necessarily correlated with EPSP amplitude at the soma[41], we normalized each spine's responses so that each spine had equal weight when computing summed spine inputs. Summed spine inputs were computed as the average spine response to each stimulus, bootstrapped 100 times. We also compared tuning of populations of spine inputs to somatic output by including response amplitude in the calculation and by computing the fraction of spines that preferentially respond to the preferred direction.

DSI was computed as:

$$\frac{\text{Preferred} - \text{Null}}{\text{Preferred} + \text{Null}}$$

Whole-cell recording. Membrane potential recordings were median filtered with a 30 to 100 sample window to remove action potentials and binned to 5 ms. Responses to individual stimulus cycles were extracted for V_m and spikes separately. Mean (F0) and modulation amplitudes (F1 and F2) of each cycle were computed via Fast Fourier Transform (MATLAB). V_m and spiking peak responses were computed as previously described[14]. Some cells exhibited V_m modulation at F2, so we also included the F2 component when computing V_m responses. For computing V_m noise, we aligned cycle responses across trials, then took the standard deviation for each time point. V_m standard deviation was computed as the mean of this standard deviation value for each stimulus.

Conductance measurements were made in current-clamp mode[11,14]. Multiple current steps depolarized or hyperpolarized the neuron close to the reversal potential for inhibition and excitation, respectively. Leak-subtracted synaptic conductances were computed by estimating G_{leak} using the blank stimulus and then performing a linear fit of measured membrane potential responses at different current injections. Mean and standard deviation of synaptic conductances were computed with a bootstrap (100 iterations). Cells were excluded from further analysis if negative conductances were extracted across multiple stimuli. To predict membrane potential responses from empirically measured synaptic conductances, we computed stimulus-dependent responses as previously described[22]:

$$\Delta V(\theta) = \frac{g_L R_L + g_e(\theta) R_e + g_i(\theta) R_i}{g_L + g_e(\theta) + g_i(\theta)} - V_r$$

in which R_L is –50 mV, V_r is –50 mV, R_e is 0 mV, R_i is –70 mV, θ is the direction of motion, g_L is the measured leak conductance, and $g_e(\theta)$ and $g_i(\theta)$ are measured synaptic conductances.

Optogenetic stimulation experiments compared visually evoked subthreshold responses under blue light stimulation with the responses obtained without blue light stimulation. Modulation ratio was computed as the response amplitude with blue light on divided by the response amplitude with blue light off.

For connectivity mapping, membrane potential traces were median filtered with a time window of 1.2 ms. We defined the prestimulus membrane potential as the membrane potential in the 9 ms before IPSP onset. Because of spontaneous activity in vivo, single trials were excluded if cells showed large depolarizations (>5 mV) relative to the prestimulus membrane potential. Significant IPSPs were defined as average IPSPs exceeding three standard deviations below the mean of the prestimulus membrane potential. We used the centroid of the significant IPSP field for distance measurements from single cells. Ellipse fits of the binarized significant IPSP array were computed using the MATLAB function 'regionprops.m'.

Intrinsic signal imaging. Single-condition maps were computed by comparing whether reflectance changes evoked by a single-stimulus condition could be discriminated from reflectance changes evoked across all presented stimuli[42]. To discriminate a single-condition stimulus at each pixel, reflectance changes across all stimuli were combined into a normalized histogram, and then a pixel's single condition response was computed non-parametrically as the probability of the area under a ROC curve (using the trapezoidal rule). Maps were filtered as previously described using a bandpass fermi filter[12]. Single bouton direction preferences were compared to the direction preference of the intrinsic signal direction preference map contained within the 100 μm two photon field of view. Somatic direction preference was compared to the direction preference of the intrinsic map at the location of the cell. For STOMPM, stimulation grids were aligned to blood vessel reference maps for intrinsic signal imaging using an affine transform. We computed binary masks for each stimulation spot, and used these masks to measure intrinsic signal direction preference at single stimulation spots.

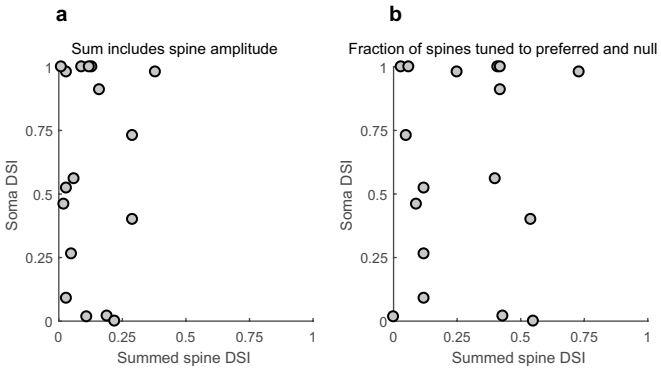
Statistics. Sample sizes are similar to others used in the field. No statistical methods were used to predetermine sample size. Inclusion criteria for each experiment are detailed for methods. The experiments were not randomized. The experimenter was blind to location in the direction preference map when performing map-related experiments; otherwise, the investigators were not blinded to allocation during experiments and outcome assessment. To test whether two distributions of direction preference were significantly different from random, we compared the median difference with a null distribution generated from Monte Carlo simulations ($n = 1{,}000$). For each Monte Carlo simulation, we calculated the median difference between two randomly sampled distributions of direction preferences drawn from a uniform distribution ranging from 0° to 359° with sample sizes equivalent to the measured distributions. Statistical tests were non-parametric and two-sided, except for the Monte Carlo significance tests, which were one-sided. All correlations values reported were computed using Spearman's correlation.

Code availability. Analyses were performed using MATLAB using standard functions. Custom code is available from the corresponding author upon reasonable request.

Data availability. Source data are provided for graphical data representations in Figs. 1d–f, 2c–g, 3c–f, h–j, and 4f, k, m, n. Data are available from the corresponding author upon reasonable request.

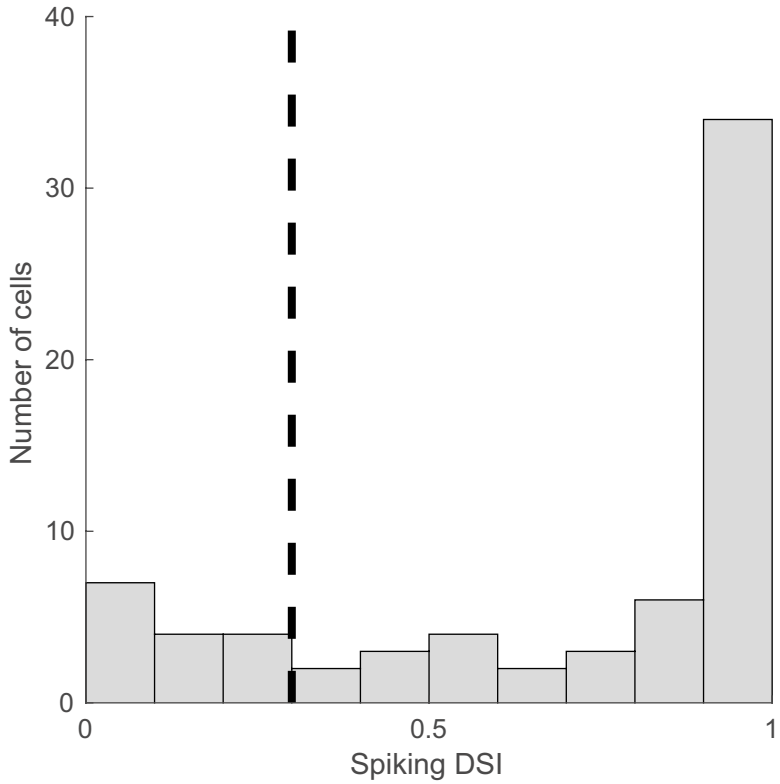
33. Dimidschstein, J. et al. A viral strategy for targeting and manipulating interneurons across vertebrate species. *Nat. Neurosci.* **19**, 1743–1749 (2016).
34. Govorunova, E. G., Sineshchekov, O. A., Janz, R., Liu, X. & Spudich, J. L. Natural light-gated anion channels: A family of microbial rhodopsins for advanced optogenetics. *Science* **349**, 647–650 (2015).
35. Peirce, J. W. PsychoPy—Psychophysics software in Python. *J. Neurosci. Methods* **162**, 8–13 (2007).
36. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
37. Baker, C. A., Elyada, Y. M., Parra, A. & Bolton, M. M. Cellular resolution circuit mapping with temporal-focused excitation of soma-targeted channelrhodopsin. *eLife* **5**, e14193 (2016).
38. Wu, C., Ivanova, E., Zhang, Y. & Pan, Z. H. rAAV-mediated subcellular targeting of optogenetic tools in retinal ganglion cells *in vivo*. *PLoS ONE* **8**, e66332 (2013).
39. Pnevmatikakis, E. A. & Giovannucci, A. NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. *J. Neurosci. Methods* **291**, 83–94 (2017).
40. Sage, D., Prodanov, D., Tinevez, J. Y. & Schindelin, J. in *ImageJ User & Developer Conference* (Luxembourg, 2012).
41. Sobczyk, A., Scheuss, V. & Svoboda, K. NMDA receptor subunit-dependent [Ca^{2+}] signaling in individual hippocampal dendritic spines. *J. Neurosci.* **25**, 6037–6046 (2005).
42. Purushothaman, G., Khaytin, I. & Casagrande, V. A. Quantification of optical images of cortical responses for inferring functional maps. *J. Neurophysiol.* **101**, 2708–2724 (2009).

© 2018 Springer Nature Limited. All rights reserved.



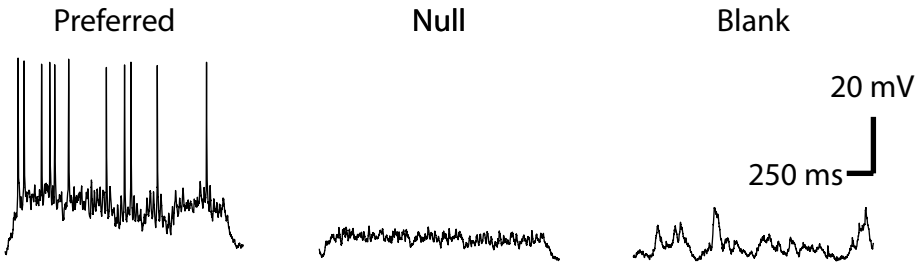
Extended Data Fig. 1 | Summed spine inputs fail to predict somatic direction selectivity, regardless of the method used to compute the sum. a, No significant correlation between the DSI of summed spine inputs (with amplitude included) and somatic DSI. Spearman's $r = -0.11$, $P = 0.68$, $n = 17$. **b**, No significant correlation between the fraction of spines that respond more strongly to the preferred direction and somatic DSI. Spearman's $r = -0.082$, $P = 0.75$, $n = 17$.

© 2018 Springer Nature Limited. All rights reserved.



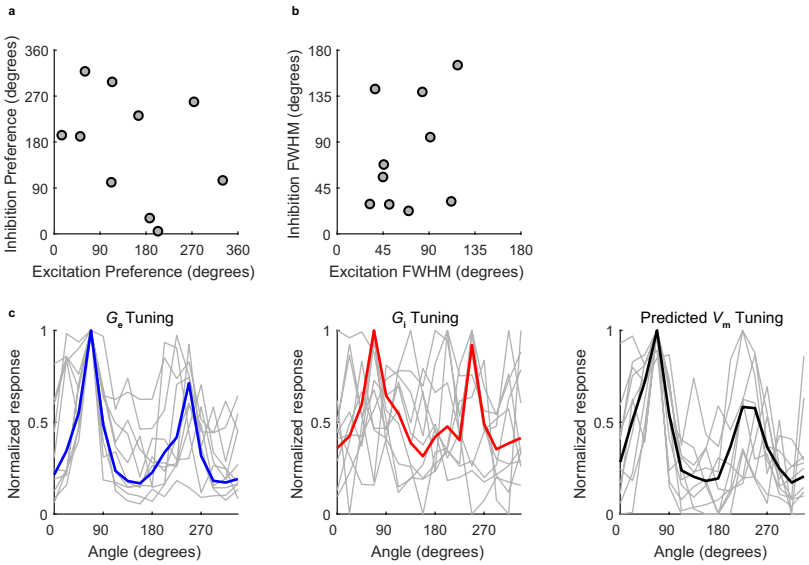
Extended Data Fig. 2 | Distribution of spiking DSI. Dashed line indicates cutoff of DSI > 0.3; $n = 69$ cells with spiking responses.

© 2018 Springer Nature Limited. All rights reserved.



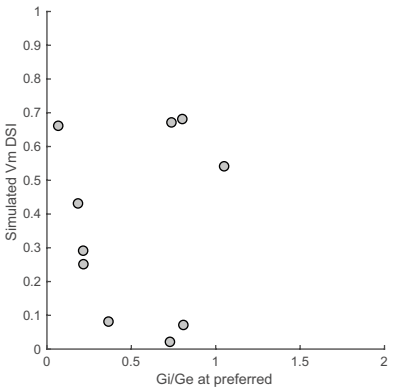
Extended Data Fig. 3 | Example of noise suppression at null stimulus relative to blank. Figure shows responses to preferred, null and blank.

© 2018 Springer Nature Limited. All rights reserved.



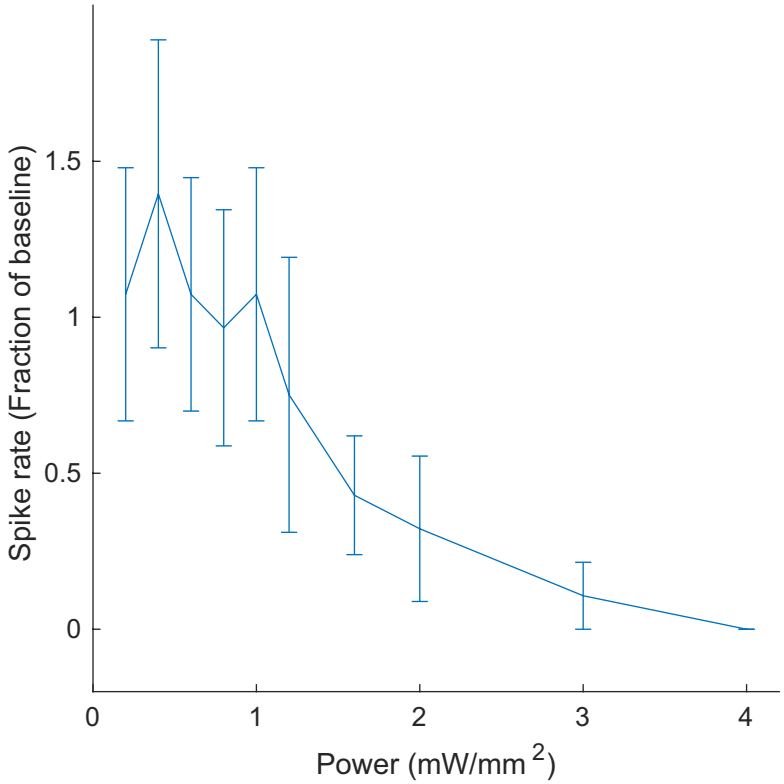
Extended Data Fig. 4 | Direction tuning fits for excitatory and inhibitory conductances. a, Difference in direction preference of excitation and inhibition are significantly greater than chance; Monte Carlo significance test, $P = 0.023$; difference in direction preference, $135° \pm 95°$, median $\pm$ IQR, $n = 10$ cells from 7 animals. **b**, FWHM of excitation and inhibition were not significantly different. FWHM $61° \pm 46°$ and $61° \pm 110°$ for excitation and inhibition, respectively; median $\pm$ IQR, $n = 10$, Wilcoxon sign-rank $P = 0.70$. **c**, Individual (grey) and population average (coloured) tuning curves for G_e, G_i and predicted V_m, peak-aligned to excitation.

© 2018 Springer Nature Limited. All rights reserved.



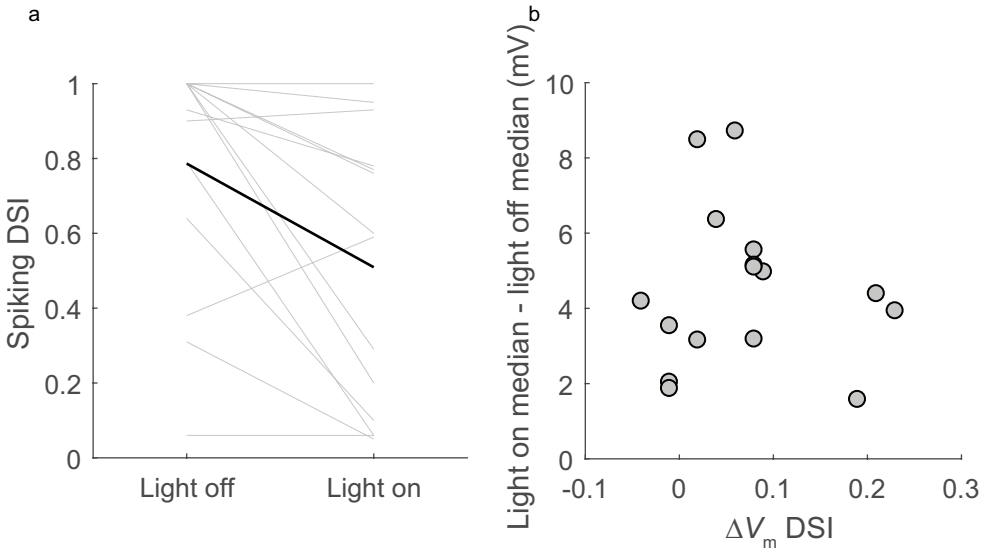
Extended Data Fig. 5 | *I*/*E* ratio at preferred direction is not correlated with simulated subthreshold direction selectivity. Spearman's $r = 0.0061$, $P = 1$, $n = 10$ cells from 7 animals.

© 2018 Springer Nature Limited. All rights reserved.



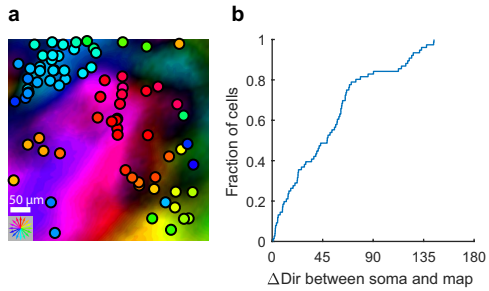
Extended Data Fig. 6 | Putative GABAergic neuron directly suppressed by blue light. Error bars, mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



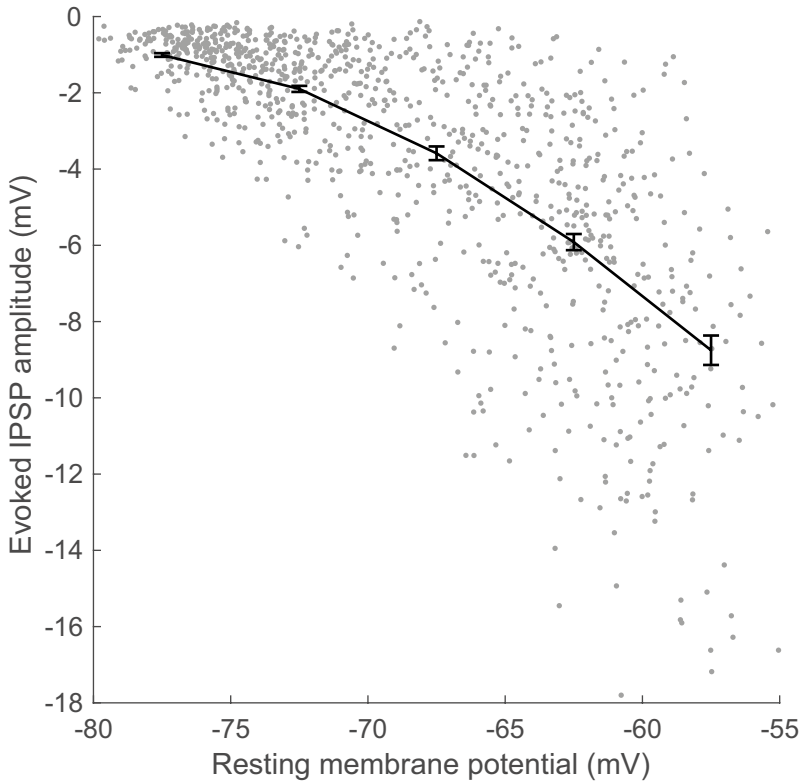
Extended Data Fig. 7 | Additional data related to blue light photoinhibition of GABAergic neurons. a, Optogenetic suppression of GABAergic neurons significantly reduces spiking direction selectivity; Wilcoxon sign-rank, $n = 14$ cells with spiking responses, $P = 0.0049$. Black line, mean; grey lines, single cells. **b**, Absolute V_m depolarization induced by blue light is not related to optogenetic changes in V_m direction selectivity (computed as the difference in DSI between light off and light on conditions); Spearman's $r = 0.11$, $P = 0.70$, $n = 14$ cells with spiking responses from 4 animals.

© 2018 Springer Nature Limited. All rights reserved.



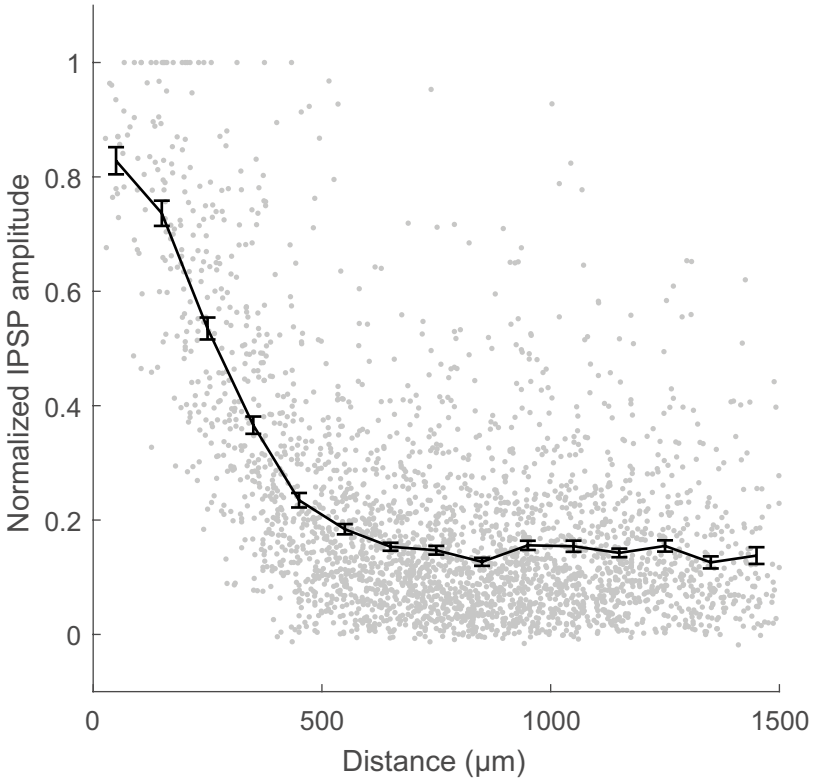
Extended Data Fig. 8 | Alignment of GABAergic neurons with intrinsic signal polar direction map. a, Underlying intrinsic signal polar direction map with direction-tuned GABAergic neurons overlaid. **b**, Direction preferences of inhibitory neurons and intrinsic signal direction preference map are significantly more similar than chance; $P < 0.001$, Monte Carlo significance test, $n = 76$ direction-selective neurons from 3 planes in 1 animal.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Reversal potential of optogenetically evoked PSPs is consistent with inhibition. Grey points are individual data points; black is mean ± s.e.m. Data come from individual stimulation trials from one cell.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Relationship of IPSP amplitude and distance. Grey points are individual data points; black is binned mean ± s.e.m. Data come from trial-averaged stimulation responses from $n = 21$ cells from 7 animals.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0341-6

Pancreatic islets communicate with lymphoid tissues via exocytosis of insulin peptides

Xiaoxiao Wan[1], Bernd H. Zinselmeyer[1], Pavel N. Zakharov[1], Anthony N. Vomund[1], Ruth Taniguchi[2], Laura Santambrogio[3], Mark S. Anderson[2], Cheryl F. Lichti[1] & Emil R. Unanue[1]*

Tissue-specific autoimmunity occurs when selected antigens presented by susceptible alleles of the major histocompatibility complex are recognized by T cells. However, the reason why certain specific self-antigens dominate the response and are indispensable for triggering autoreactivity is unclear. Spontaneous presentation of insulin is essential for initiating autoimmune type 1 diabetes in non-obese diabetic mice[1,2]. A major set of pathogenic CD4 T cells specifically recognizes the 12–20 segment of the insulin B-chain (B:12–20), an epitope that is generated from direct presentation of insulin peptides by antigen-presenting cells[3,4]. These T cells do not respond to antigen-presenting cells that have taken up insulin that, after processing, leads to presentation of a different segment representing a one-residue shift, B:13–21[4]. CD4 T cells that recognize B:12–20 escape negative selection in the thymus and cause diabetes, whereas those that recognize B:13–21 have only a minor role in autoimmunity[3–5]. Although presentation of B:12–20 is evident in the islets[3,6], insulin-specific germinal centres can be formed in various lymphoid tissues, suggesting that insulin presentation is widespread[7,8]. Here we use live imaging to document the distribution of insulin recognition by CD4 T cells throughout various lymph nodes. Furthermore, we identify catabolized insulin peptide fragments containing defined pathogenic epitopes in β-cell granules from mice and humans. Upon glucose challenge, these fragments are released into the circulation and are recognized by CD4 T cells, leading to an activation state that results in transcriptional reprogramming and enhanced diabetogenicity. Therefore, a tissue such as pancreatic islets, by releasing catabolized products, imposes a constant threat to self-tolerance. These findings reveal a self-recognition pathway underlying a primary autoantigen and provide a foundation for assessing antigenic targets that precipitate pathogenic outcomes by systemically sensitizing lymphoid tissues.

On the basis of previous studies demonstrating constrained T-cell migration during limited antigen recognition[9–12], we investigated insulin presentation in peripheral lymph nodes by two-photon microscopy of lymph-node explants following transfer of insulin-specific T cells (Fig. 1a). These were transferred together with wild-type CD4 T cells as a control; each population of transferred cells was labelled with a different fluorescent probe (Fig. 1b). We monitored the individual trajectories of transplanted T cells within the same region of the lymph nodes and quantified their motility (Extended Data Fig. 1a). We first performed this assay with a control CD4 T cell (10E11), which recognizes hen egg lysozyme (HEL). These experiments confirm that limited antigen recognition that is insufficient to trigger cell division can be detected by a decrease in the mean velocity of T cells (Fig. 1c, Extended Data Fig. 1b, Supplementary Video 1).

Widespread presentation of insulin peptides was demonstrated by reduced motility of the B:12–20-specific 8F10 T cells in the pancreatic (pLN), inguinal (iLN), mesenteric (mLN) and axillary (aLN) lymph nodes of non-obese diabetic (NOD) mice, relative to wild-type CD4 T cells (Fig. 1d, Supplementary Video 2). Motility was reduced to a similar degree on day 1 or day 5 of imaging (Extended Data Fig. 1c), and was unaffected by switching the labelling of the fluorescent probes (Extended Data Fig. 1d). The diffuse, rather than clustered, pattern of motility arrest indicates that presentation of insulin peptides was limiting and was not restricted to selected antigen-presenting cells (APCs). Motility of 8F10 T cells was also reduced in μMT and Batf3$^{-/-}$ mice, which are deficient in B cells and XCR1^{+} dendritic cells, respectively (Extended Data Fig. 1e).

We performed three experiments to interrogate key parameters of antigen recognition by 8F10 T cells. First, we examined B16A mice, which are deficient in both *Ins1* and *Ins2* but express a proinsulin transgene with a Tyr16Ala substitution in the B chain[1]. This mutant insulin is bioactive but is not immunogenic to B:12–20- or B:13–21-specific T cells. There was no reduction in motility of 8F10 T cells in the B16A mouse recipients, demonstrating that the effects on T cell motility require specific epitope recognition by the 8F10 T cells (Fig. 1e, Supplementary Video 3). Second, we investigated whether prior recirculation through the pLN was required for insulin recognition in other sites. Surgical removal of pLNs did not influence the motility arrest of 8F10 T cells in the iLNs (Fig. 1f). Third, we detected motility arrest of 8F10 T cells in diabetes-resistant B6 mice harbouring the I-A^{g7} haplotype (B6g7) (Fig. 1g) but not in NOD mice with the H2b haplotype (Extended Data Fig. 1f). Therefore, peripheral insulin presentation to 8F10 T cells requires I-A^{g7} and is not restricted to the NOD strain.

The motility of 4F7 T cells, which specifically recognize the B:13–21 epitope, was also markedly reduced in the pLNs and iLNs of NOD recipients (Extended Data Fig. 1g). By contrast, the 8.3 CD8 T cells, which recognize the islet-specific glucose-6-phosphatase-related protein (IGRP[13], a protein that is expressed in the endoplasmic reticulum of β-cells), exhibited reduced motility in the pLN but not in the iLN (Extended Data Fig. 1h). Therefore, epitopes of insulin, but not those from IGRP, a cell-associated antigen, are systemically available.

We hypothesized that presentation of the low concentrations of circulating insulin (about 40 pM) might require insulin receptor-mediated uptake by APCs. To test this, we examined the effects of S961, an insulin receptor antagonist[14]. In assays on cultured cells, S961 impaired the ability of concanavalin A (ConA)-activated macrophages to present insulin (Extended Data Fig. 1i). In vivo blockade of insulin receptor by infusion of mice with S961 via osmotic pump caused a sustained increase in blood glucose levels (Extended Data Fig. 1j), permitting two-photon microscopy analysis (Extended Data Fig. 1k). A significant reduction in mean velocity of transferred 4F7 T cells was observed in lymph nodes of control mice infused with phosphate-buffered saline (PBS) (Fig. 1h, Supplementary Video 4). Although the motility of 4F7 T cells was also arrested after S961 infusion, the magnitude of the reduction was significantly smaller than with PBS (Fig. 1h, Supplementary Video 4). Therefore, blockade of insulin receptor-mediated uptake of insulin partially abrogated recognition by the 4F7 T cells, suggesting that free insulin peptides are an additional source of the B:13–21

[1]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA. [2]Diabetes Center, University of California, San Francisco, San Francisco, CA, USA. [3]Department of Pathology, Albert Einstein College of Medicine, New York, NY, USA. *e-mail: unanue@wustl.edu

© 2018 Springer Nature Limited. All rights reserved.

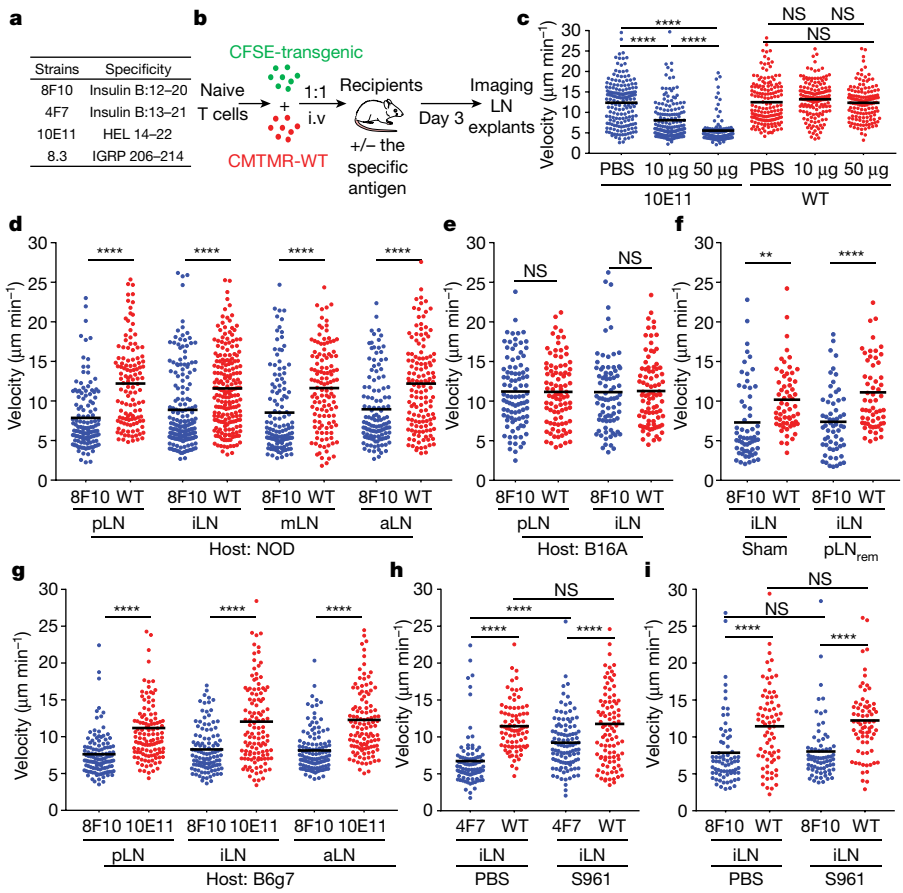


Fig. 1 | Peripheral insulin presentation is systemic, epitope-specific, and occurs physiologically. a, Specificities of the antigen-specific T cells. **b**, The scheme of the transplant and two-photon imaging experiments using CFSE and CMTMR labels for transgenic and wild-type (WT) T cells, respectively. LN, lymph node. **c–i**, Mean track velocities of: 10E11 and wild-type (WT) CD4 T cells in mice given PBS or the indicated amounts of HEL (**c**); 8F10 and wild-type CD4 T cells in NOD (**d**) or B16A (**e**) mice; 8F10 and wild-type CD4 T cells in NOD mice after surgical removal of the pancreatic lymph nodes (pLN_{rem}) or sham surgery (**f**); 8F10 and 10E11 CD4 T cells in B6g7 mice 24 h after transfer (**g**); 4F7 and wild-type (**h**) or 8F10 and wild-type (**i**) CD4 T cells in NOD mice infused with S961 or PBS. pLN, pancreatic lymph node; iLN, inguinal lymph node; mLN, mesenteric lymph node; aLN, axillary lymph node. Data are pooled results from at least three independent experiments. Each dot represents the velocity of one T cell, and the bar denotes the mean of all T cells in the group. $^{**}P < 0.001$; $^{****}P < 0.0001$; one-way ANOVA with Sidak's multiple comparisons test. NS, not significant.

epitope. By contrast, the motility arrest of 8F10 T cells remained at a comparable level in recipients infused with PBS or S961 (Fig. 1i, Supplementary Video 5), indicating that the presence of B:12–20 is independent of insulin receptor-mediated uptake of insulin. This epitope must therefore derive from insulin peptides that reach the peripheral lymphoid organs. Of note, APCs expressing autoimmune

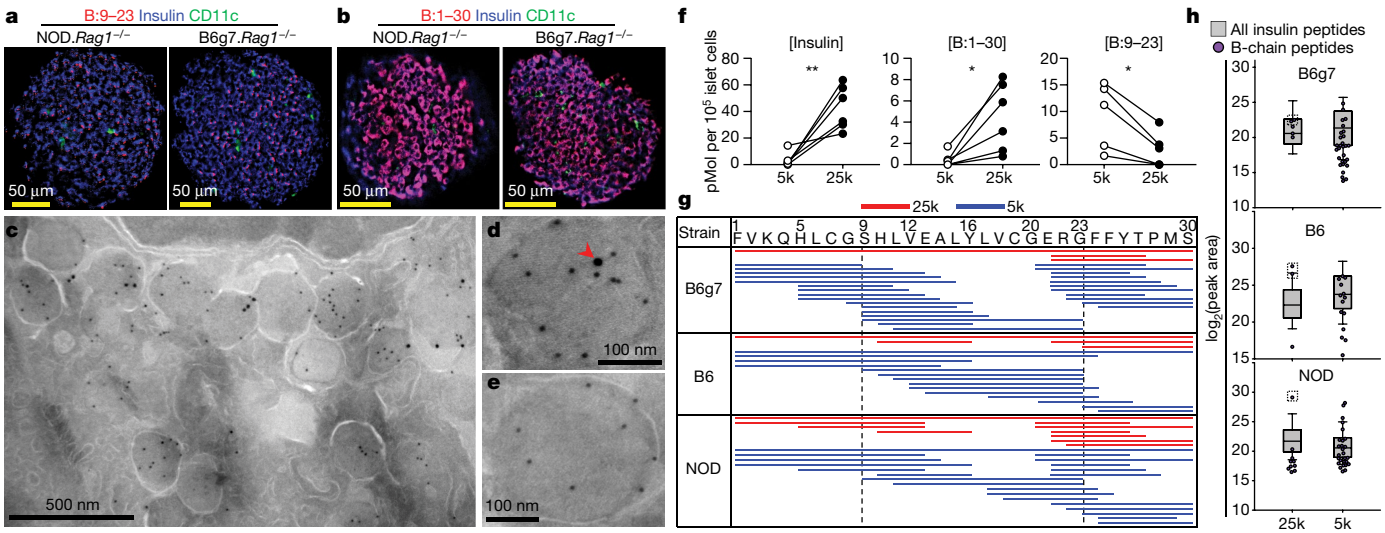
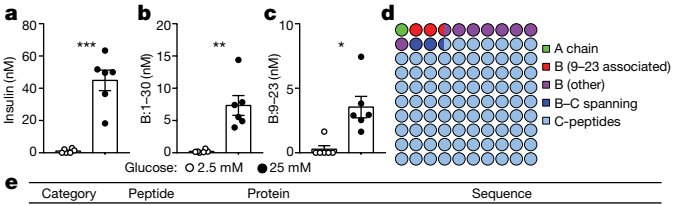


Fig. 2 | Generation of insulin peptides in β-cell granules. **a**, **b**, Immunofluorescence of isolated islets stained for B:9–23 (**a**) or B:1–30 (**b**), CD11c and insulin. Data are representative of 50 islets per group in three independent experiments. **c**, Immunogold electron microscopy showing antibodies against B:1–30 (large gold) and insulin (small gold) in a representative β-cell. **d**, **e**, A representative granule that contains both B:1–30 and insulin (**d**) or insulin only (**e**) is shown. The arrowhead in **d** indicates the B:1–30 peptide. Data are representative of 317 granules analysed in three independent experiments. **f**, Competitive ELISA showing quantification of insulin, B:1–30 and B:9–23 in granules isolated by centrifugation of islets from B6g7 mice at 5,000*g* (5k) or 25,000*g* (25k). Each line represents one paired experiment using 4–8 mice. $^{*}P < 0.05$; $^{**}P < 0.01$; two-tailed paired Student's *t*-test. **g**, Peptide coverage of insulin B chain by sequences identified in 25k (red) and 5k (blue) β-cell granules using nLC–MS/MS analysis. Each line represents the alignment of individual peptides with insulin-2 B:1–30. Data are from four independent analyses using islets from 8–10 mice per strain. **h**, Box plot of $\log_2$(mass spectrometry peak area), showing the abundance of individual insulin B-chain peptides (purple) in the 25k and 5k granules relative to all insulin peptides, including the C-peptides (box). Boxes with dashed outlines denote B:1–30 with a high abundance. Box plots show the median, box edges represent the first and third quartiles, and the whiskers extend to 1.5× interquartile range.

© 2018 Springer Nature Limited. All rights reserved.



e

Category	Peptide	Protein	Sequence
B (9–23 associated)	B:1–30	Insulin-1	FVKQHLCGPHLVEALYLVCGERGFFYTPKS
	B:1–30	Insulin-2	FVKQHLCGSHLVEALYLVCGERGFFYTPMS
	B:9–23	Insulin-2	SHLVEALYLVCGERG
	B:11–23	Insulin-1,2	LVEALYLVCGERG
	B:11–30	Insulin-1	LVEALYLVCGERGFFYTPKS
	B:12–22	Insulin-1,2	VEALYLVCGER
	B:12–23	Insulin-1,2	VEALYLVCGERG
	B:13–30	Insulin-1	EALYLVCGERGFFYTPKS
	B:15–23	Insulin-1,2	LYLVCGERG
A chain	A:12–21	Insulin-1,2	SLYQLENYCN
	A:14–20	Insulin-1,2	YQLENYC
	A:14–21	Insulin-1,2	YQLENYCN
B–C spanning	B25–C23	Insulin-1	**FYTPKS**RREVEDPQVEQLELGGSPGDLQTLA
	B26–C21	Insulin-1	**YTPKS**RREVEDPQVEQLELGGSPGDLQT
	B26–C22	Insulin-2	**YTPMS**RREVEDPQVAQLELGGGPGAGDLQ
	B26–C23	Insulin-1	**YTPKS**RREVEDPQVEQLELGGSPGDLQTLA
	B26–C23	Insulin-2	**YTPMS**RREVEDPQVAQLELGGGPGAGDLQT
	B26–C25	Insulin-2	**YTPMS**RREVEDPQVAQLELGGGPGAGDLQTLA

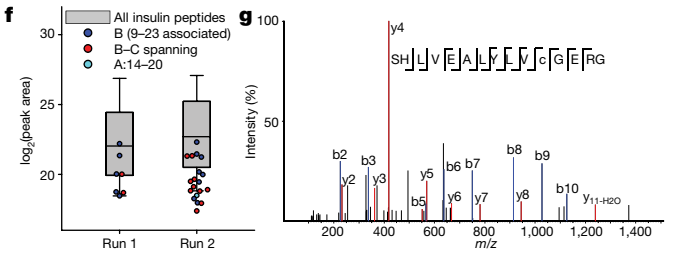


Fig. 3 | Secretion of insulin peptides into the circulation upon glucose stimulation. a–**c**, Insulin (**a**), B:1–30 (**b**) and B:9–23 (**c**) secreted from islets of B6g7 mice during one hour after stimulation with 2.5 mM or 25 mM glucose, quantified by competitive ELISA. Each point represents an independent experiment. $*P < 0.05$; $**P < 0.01$; $***P < 0.005$; two-tailed paired Student's *t*-test. **d**, A 10 × 10 dot plot representing the coverage of insulin peptide sequences identified by nLC–MS/MS in supernatants of cultured islets stimulated with 25 mM glucose. Each dot represents 1% coverage of the total. **e**, Summary of selected insulin peptides containing defined immunogenic epitopes. The following epitopes are underlined: B:12–20 (red), B:13–21 (blue), B:15–23 (green) and A:14–20 (black). In B–C-spanning peptides, the residues of the B chain are shown in bold. **f**, Box plot of $\log_2$(peak area) showing abundance of individual B:9–23-associated peptides (blue), B–C-spanning peptides (red) and the A:14–20 peptide (cyan) relative to all insulin peptides (box). Box plots show the median, box edges represent the first and third quartiles, and the whiskers extend to 1.5× interquartile range. **g**, The mass spectrum of a peptide sequence identified in mouse urine that contains all residues of the insulin B:9–23 peptide, with oxidation of cysteine to cysteic acid (c).

regulator (AIRE)[15] are not a major source of insulin peptides (Extended Data Fig. 1l).

We identified insulin peptides in β-cell granules using peptide-specific monoclonal antibodies and mass spectrometry analysis. We used the monoclonal antibody AIP, which is specific for B:9–23[3,6], and generated a new monoclonal antibody (clone 6F3.B8) by immunization with the entire insulin B-chain (B:1–30). The two antibodies were not cross-reactive and neither recognized native insulin (Extended Data Fig. 2a–d). Notably, presentation of B:1–30 activated insulin-reactive T cells without the need for internal processing (Extended Data Fig. 2e).

We previously identified B:9–23 in a set of LAMP1-positive vesicles in β-cells[3]. These vesicles are distinct from insulin-containing dense core granules and can be separated from them by differential centrifugation[6]. They are consistent with the crinophagic bodies that result from fusion of the dense core granules to lysosomes[16–18], and contain peptides that preferentially stimulate 8F10 T cells[6]. Immunofluorescence with AIP showed a punctate pattern of B:9–23 staining in β-cells from NOD.*Rag1*$^{-/-}$ or B6g7.*Rag1*$^{-/-}$ mice (Fig. 2a). By contrast, B:1–30 staining using 6F3.B8 was more diffuse in nearly all the β-cells and co-stained with insulin (Fig. 2b). Using double immunogold-labelling antibodies, we detected B:1–30 in granules containing insulin (Fig. 2c). Many granules (106 out of 317, 33%) contained both B:1–30 and insulin (Fig. 2d), and the rest contained only insulin (Fig. 2e). AIP did not stain islets satisfactorily after labelling with immunogold.

Regular secretory granules obtained by centrifugation at 25,000*g* (25k) contained significantly higher amounts of insulin than the fraction obtained at 5,000*g* (5k), which includes the crinophagic bodies (Fig. 2f). B:1–30 was primarily found in the 25k fraction and not in the 5k fraction, but in concentrations about one tenth that of insulin (Fig. 2f). By contrast, B:9–23 was significantly more abundant in the granules in the 5k fraction (Fig. 2f).

We analysed the peptidomes of granules prepared from B6g7 mice, B6 mice or 3-week old female NOD mice, by nanoflow liquid chromatography–tandem mass spectrometry (nLC–MS/MS). In all three strains, granules from the 25k fraction mostly contained sequences from the insulin C-peptide, the intact B:1–30, and a few small peptides from the B chain (Fig. 2g, h, Supplementary Table 1, Extended Data Fig. 3a). By contrast, the granules from the 5k fraction, contained more diverse short sequences from throughout the B chain (Fig. 2g, h, Supplementary Table 1). Peptides derived from the 9–23 region, such as B:9–23 and B:11–23 (Extended Data Fig. 3b), were identified exclusively in the 5k granules of all three mouse strains. Manual interrogation of unassigned spectra only identified two putative hybrid peptides in the 5k granules (Extended Data Fig. 3c), a C-peptide–islet amyloid polypeptide (IAPP) fusion, and a fusion of the N terminus of the C-peptide of insulin-2 and the C terminus of the C-peptide of insulin-1. Peptides from other proteins were present at much lower levels in comparison to those from insulin.

Examination of β-cell granules from human islets revealed a striking similarity in the segregation of peptides between 5k and 25k fractions to that from the mouse islets (Extended Data Fig. 4a, Supplementary Table 1). The human 25k granules contained the intact B chain and a limited number of short sequences. The 5k fraction contained many short peptides, including a sequence representing B:11–30 (Extended Data Fig. 4b), containing the HLA(DQ8)-binding B:11–23 determinant, which is recognized by peripheral T cells in patients with type 1 diabetes[19].

Islets stimulated with 25 mM glucose secreted insulin (Fig. 3a) along with lower concentrations of peptides that were recognized by 6F3.B8 (Fig. 3b) or AIP (Fig. 3c). Secretion of insulin or insulin peptides was not affected when glucose challenge was carried out in the presence of protease inhibitors (Extended Data Fig. 5a), indicating that the peptides were not generated extracellularly.

We characterized insulin peptides secreted by β-cells using nLC–MS/MS. Most of these peptides were derived from the C-peptide, along with B-chain-derived sequences related to the 9–23 region and spanning the B-chain–C-peptide (B–C) junction (Fig. 3d, Supplementary Table 2). Many of the peptides were identical to or contained pathogenic epitopes that were identified using diabetogenic T cells as probes[1,3,4,20–22] (Fig. 3e, Extended Data Fig. 5b–e). The intact B chain contained identical sequences to peptides identified in the 25k granules, whereas B:9–23 and B:11–23 were identical to peptides in the 5k granules (Fig. 3e, Supplementary Table 2). Synthetic versions of peptides associated with B:9–23 activated T cells specific for B:12–20 as well as those specific for B:13–21 (Extended Data Fig. 6). In general, these potentially immunogenic peptides were present at low relative abundance (Fig. 3f, Supplementary Table 2).

We identified a form of B:9–23 containing cysteine oxidized to cysteic acid in mouse urine using antibody capture (Fig. 3g); this is a modification that can occur during sample preparation (Extended Data Fig. 7a). This finding indicates that B:9–23 is present in the circulation. Indeed, fluorochrome-labelled B:9–23 was rapidly displayed

© 2018 Springer Nature Limited. All rights reserved.

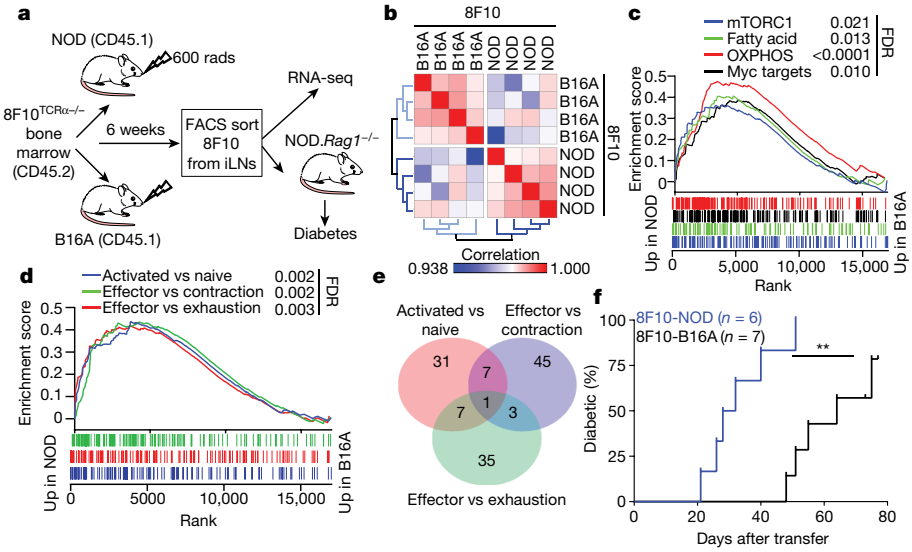


Fig. 4 | Acquisition of an effector-like phenotype by 8F10 T cells during antigen recognition. a, Experimental design for **b**–**f**. **b**, Pearson's correlation matrix showing hierarchical clustering of RNA-seq of 8F10 T cells sourced from NOD or B16A hosts. **c**, GSEA enrichment plots showing a significant correlation (determined by false discovery rate (FDR) $Q < 0.05$) of genes upregulated in the 8F10-NOD samples with four hallmark datasets associated with metabolism pathways. **d**, GSEA enrichment plots showing a significant correlation of genes upregulated in the 8F10-NOD samples with three immunological signature datasets depicting T cell activation and effector function. **e**, A Venn diagram showing the number of overlapping genes among the three gene sets in **d**. **f**, Incidence of diabetes in NOD.*Rag1*$^{-/-}$ recipients adoptively transferred with 8F10 T cells isolated from the iLNs of NOD or B16A mice six weeks after bone marrow transfer. $^{**}P < 0.005$; log-rank test. Data represent cumulative results of three independent transfers.

by I-A^{g7}-expressing APCs in spleen but not in thymus following intravenous injection (Extended Data Fig. 7b, c).

The widespread presentation of insulin peptides in lymphoid tissues influences the biology of T cells. We generated a bone marrow chimaera model in which we transferred a small number of bone marrow stem cells from CD45.2 8F10 mice deficient in T-cell receptor alpha chain (8F10$^{TCR\alpha-/-}$) into non-lethally irradiated NOD or B16A hosts (CD45.1) (Fig. 4a). This resulted in the development of a small number of 8F10 T cells (0.5–2%) among the endogenous CD4 T-cell repertoire (Extended Data Fig. 8a). We performed RNA sequencing analysis (RNA-seq) on isolated 8F10 T cells from iLNs of both hosts (Fig. 4a).

Hierarchical clustering using Pearson's correlation revealed differences between the transcriptomes of 8F10 T cells sourced from NOD (8F10-NOD) and B16A (8F10-B16A) hosts (Fig. 4b). Gene-set enrichment analysis (GSEA) showed significant correlations between transcripts that were upregulated in the 8F10-NOD T cells with biological pathways involving oxidative phosphorylation (OXPHOS), Myc targets, fatty acid metabolism, mTOR complex 1 (mTORC1) signalling and DNA repair (Extended Data Fig. 8b). The four most highly ranked gene sets (Fig. 4c) were associated with metabolic pathways, and involved transcripts encoding key kinases, intermediates and transcription factors (Extended Data Fig. 8c) that have been shown to support T cell proliferation and functions[23].

This metabolic reprogramming in 8F10 T cells from NOD mice was associated with an effector-like phenotype (Fig. 4d). The gene sets that were upregulated in these cells are also highly expressed in CD4 T cells upon stimulation[24], in CD8 T cells at the peak of expansion in comparison to the contraction phase[25], and in CD8 effectors in comparison to exhausted T cells[26]. There is little overlap among these three sets of transcripts (Extended Data Fig. 9a, Fig. 4e). According to GSEA, neither T cell set correlated with anergic CD4[27] or tolerant CD8 T cells[28] (Extended Data Fig. 9b).

Functional analysis at the six-week time point revealed a higher capacity of effector cytokine (TNF and IFNγ) production (Extended Data Fig. 10a) and cell proliferation (Extended Data Fig. 10b) in 8F10-NOD T cells. Neither T cell set expressed molecules associated with anergy and exhaustion[26,27,29] (Extended Data Fig. 10c). Similar results were obtained for T cells analysed nine weeks after the bone marrow transfer (Extended Data Fig. 10d–f). Of note, when the two sets of T cells were transferred into NOD.*Rag1*$^{-/-}$ recipients, the onset of diabetes was accelerated by the 8F10-NOD set (Fig. 4f). Therefore, 8F10 T cells acquired an effector-like phenotype during peripheral antigen recognition, supported by transcriptional reprogramming and increased diabetogenicity.

In summary, peptide exocytosis is a normal response of β-cells that represents a mechanism of communication with the lymphoid tissues. Similar mechanisms may apply to other endocrine organs that also contain crinophagic granules. Examining the released peptides may enable better-targeted identification of T cell responses; a set of responses that could be extensive, given the diversity of exocytosed moieties. Previous studies have shown that ablation of all lymph nodes eradicates the pathogenic T cell repertoire and abolishes diabetes[30], emphasizing the importance of the entire lymphatic system in interactions with T cells. Finally, the biological outcomes described here for 8F10 T cells may vary for other insulin-reactive T cells with divergent TCR affinities. Comprehensive understanding of these outcomes will require analysis of the entire insulin-reactive T cell pool at different stages of the disease.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0341-6.

Received: 14 December 2017; Accepted: 4 June 2018;
Published online 18 July 2018.

1. Nakayama, M. et al. Prime role for an insulin epitope in the development of type 1 diabetes in NOD mice. *Nature* **435**, 220–223 (2005).
2. Unanue, E. R. Antigen presentation in the autoimmune diabetes of the NOD mouse. *Annu. Rev. Immunol.* **32**, 579–608 (2014).
3. Mohan, J. F. et al. Unique autoreactive T cells recognize insulin peptides generated within the islets of Langerhans in autoimmune diabetes. *Nat. Immunol.* **11**, 350–354 (2010).

© 2018 Springer Nature Limited. All rights reserved.

4. Mohan, J. F., Petzold, S. J. & Unanue, E. R. Register shifting of an insulin peptide–MHC complex allows diabetogenic T cells to escape thymic deletion. *J. Exp. Med.* **208**, 2375–2383 (2011).
5. Mohan, J. F., Calderon, B., Anderson, M. S. & Unanue, E. R. Pathogenic CD4$^+$ T cells recognizing an unstable peptide of insulin are directly recruited into islets bypassing local lymph nodes. *J. Exp. Med.* **210**, 2403–2414 (2013).
6. Vomund, A. N. et al. Beta cells transfer vesicles containing insulin to phagocytes for presentation to T cells. *Proc. Natl Acad. Sci. USA* **112**, E5496–E5502 (2015).
7. Wan, X., Thomas, J. W. & Unanue, E. R. Class-switched anti-insulin antibodies originate from unconventional antigen presentation in multiple lymphoid sites. *J. Exp. Med.* **213**, 967–978 (2016).
8. Wan, X. & Unanue, E. R. Unique features in the presentation of insulin epitopes in autoimmune diabetes: an update. *Curr. Opin. Immunol.* **46**, 30–37 (2017).
9. Egen, J. G. et al. Intravital imaging reveals limited antigen presentation and T cell effector function in mycobacterial granulomas. *Immunity* **34**, 807–819 (2011).
10. Le Borgne, M. et al. The impact of negative selection on thymocyte migration in the medulla. *Nat. Immunol.* **10**, 823–830 (2009).
11. Zinselmeyer, B. H. et al. In situ characterization of CD4$^+$ T cell behavior in mucosal and systemic lymphoid tissues during the induction of oral priming and tolerance. *J. Exp. Med.* **201**, 1815–1823 (2005).
12. Eickhoff, S. et al. Robust anti-viral immunity requires multiple distinct T cell–dendritic cell interactions. *Cell* **162**, 1322–1337 (2015).
13. Verdaguer, J. et al. Spontaneous autoimmune diabetes in monoclonal T cell nonobese diabetic mice. *J. Exp. Med.* **186**, 1663–1676 (1997).
14. Schäffer, L. et al. A novel high-affinity peptide antagonist to the insulin receptor. *Biochem. Biophys. Res. Commun.* **376**, 380–383 (2008).
15. Gardner, J. M. et al. Deletional tolerance mediated by extrathymic Aire-expressing cells. *Science* **321**, 843–847 (2008).
16. Smith, R. E. & Farquhar, M. G. Lysosome function in the regulation of the secretory process in cells of the anterior pituitary gland. *J. Cell Biol.* **31**, 319–347 (1966).
17. Halban, P. A. & Wollheim, C. B. Intracellular degradation of insulin stores by rat pancreatic islets in vitro. An alternative pathway for homeostasis of pancreatic insulin content. *J. Biol. Chem.* **255**, 6003–6006 (1980).
18. Weckman, A. et al. Autophagy in the endocrine glands. *J. Mol. Endocrinol.* **52**, R151–R163 (2014).
19. Yang, J. et al. Autoreactive T cells specific for insulin B:11–23 recognize a low-affinity peptide register in human subjects with autoimmune diabetes. *Proc. Natl Acad. Sci. USA* **111**, 14840–14845 (2014).
20. Wong, F. S. et al. Identification of an MHC class I-restricted autoantigen in type 1 diabetes by screening an organ-specific cDNA library. *Nat. Med.* **5**, 1026–1031 (1999).
21. Lamont, D. et al. Compensatory mechanisms allow undersized anchor-deficient class I MHC ligands to mediate pathogenic autoreactive T cell responses. *J. Immunol.* **193**, 2135–2146 (2014).
22. Chen, W. et al. Evidence that a peptide spanning the B–C junction of proinsulin is an early Autoantigen epitope in the pathogenesis of type 1 diabetes. *J. Immunol.* **167**, 4926–4935 (2001).
23. Buck, M. D., O'Sullivan, D. & Pearce, E. L. T cell metabolism drives immunity. *J. Exp. Med.* **212**, 1345–1360 (2015).
24. Constantinides, M. G., Picard, D., Savage, A. K. & Bendelac, A. A naive-like population of human CD1d-restricted T cells expressing intermediate levels of promyelocytic leukemia zinc finger. *J. Immunol.* **187**, 309–315 (2011).
25. Kaech, S. M., Hemby, S., Kersh, E. & Ahmed, R. Molecular and functional profiling of memory CD8 T cell differentiation. *Cell* **111**, 837–851 (2002).
26. Wherry, E. J. et al. Molecular signature of CD8+ T cell exhaustion during chronic viral infection. *Immunity* **27**, 670–684 (2007).
27. Macián, F. et al. Transcriptional mechanisms underlying lymphocyte tolerance. *Cell* **109**, 719–731 (2002).
28. Schietinger, A., Delrow, J. J., Basom, R. S., Blattman, J. N. & Greenberg, P. D. Rescued tolerant CD8 T cells are preprogrammed to reestablish the tolerant state. *Science* **335**, 723–727 (2012).
29. Kalekar, L. A. et al. CD4$^+$ T cell anergy prevents autoimmunity and generates regulatory T cell precursors. *Nat. Immunol.* **17**, 304–314 (2016).
30. Levisetti, M. G., Suri, A., Frederick, K. & Unanue, E. R. Absence of lymph nodes in *NOD* mice treated with lymphotoxin-β receptor immunoglobulin protects from diabetes. *Diabetes* **53**, 3115–3119 (2004).

Acknowledgements We thank J. Carrero, A. Shaw and F. Urano for critical support and review of the work; and B. Saunders, W. Beatty, H. Hu and K. Fredericks for technical assistance. This work was supported by National Institutes of Health grants DK058177 and AI14551. The laboratory also received generous support from the Kilo Diabetes & Vascular Research Foundation.

Reviewer information *Nature* thanks S. Barg, P. Santamaria and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions X.W. and E.R.U. conceived the project, designed the experiments and analysed the results. B.H.Z. and X.W. performed two-photon microscopy and analysed the data. A.N.V. and X.W. prepared samples for mass spectrometry. C.F.L. performed the nLC–MS/MS studies and analysed the data. A.N.V. generated monoclonal antibodies and isolated β-cell granules. X.W. and A.N.V. performed cell biology experiments. P.N.Z. and X.W. analysed the RNA-seq data. M.S.A., R.T. and L.S. provided key reagents and interpreted the results. X.W., C.F.L. and E.R.U. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0341-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0341-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to E.R.U.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Mice. NOD/ShiLtJ (NOD), NOD.129S7(B6)-Rag1^{tm1Mom}/J (NOD.Rag1$^{-/-}$), NOD.Cg-Tg(Ins2*Y16A)1EllIns1^{tm1Jja}Ins2^{tm1Jja}/GseJ (NOD.B16A), NOD.Cg-Tg(TcraTcrbNY8.3)1Pesa/DvsJ (8.3), NOD.129S2(B6)-Ighm^{tm1Cgn}/DoiJ (μMT), NOD.C-(Ptprc-D1Mit262)/WehiJ (NOD.CD45.2), NOD.B10Sn-H2b/J (NOD.H2b) and B6.NOD-(D17Mit21-D17Mit10)/LtJ (B6g7) mice were originally obtained from the Jackson Laboratory. NOD.10E11 TCR transgenic mice (TCRα: TRAV5D-4/TRAJ42; TCRβ: TRBV13-3/ TRBD2/TRBJ2-7) were generated using a previous protocol[5]. NOD.4F7 TCR transgenic mice, NOD.Aire$^{-/-}$, and NOD.TCRα$^{-/-}$ mice were generated by M.S.A. The 8F10 or 10E11 mice expressing the CD45.2 allotype were generated by intercrossing the original TCR transgenic line (CD45.1) with the NOD.CD45.2 mice, and the CD45.2.NOD.8F10 mice were further crossed with the NOD.TCRα$^{-/-}$ mice to generate the CD45.2 8F10$^{TCRα-/-}$ mice. B6.*Rag1*$^{-/-}$ mice were used to intercross with B6g7 mice to generate B6g7.*Rag1*$^{-/-}$ mice. All mice were bred and maintained under specific pathogen-free conditions in our animal facility. All experiments were approved by the Division of Comparative Medicine of Washington University School of Medicine in St. Louis (Accreditation number A3381-01).

Human pancreatic islets. De-identified human primary islets isolated from deceased donors were obtained from Prodo Laboratories. Experiments were judged to be 'not human subject research' by Washington University Human Research Protection Office (IRB ID # 201801183; Federalwide Assurance #FWA00002284). In total, islets from three donors were used: donor 1 (female, 57 years, BMI 21.35), donor 2 (female, 49 years, BMI 33), donor 3 (male, 28 years, BMI 34.7). Purity of the islets was between 85 and 98%. Islets were cultured in CMRL medium supplemented with 10% FBS and 50% L-cell conditioned medium[31] for recovery. The granules were isolated from ~1500 islets after 1–3 days of culture.

Antibodies. The following fluorescently conjugated antibodies were purchased from BioLegend: anti-B220 (RA3-6B2), anti-CD11c (N418), anti-CD4 (RM4-5), anti-CD45 (30-F11), anti-CD45.1 (A20), anti-CD45.2 (104), anti-CD8a (53-6.7), anti-F4/80 (BM8), anti-Vβ8.1/8.2 (KJ16-133.18), anti-CD44 (IM7), anti-CD62L (MEL-14), anti-CD25 (PC61.5) and anti-TNFα (MP6-XT22). Unconjugated or Alexa Fluor 647-labelled Rabbit anti-insulin monoclonal antibody(C27C9) was purchased from Cell Signaling Technology. Unconjugated mouse anti-insulin monoclonal antibody (E11D7) was purchased from Millipore. Alexa Fluor 594 F(ab)$_2$ donkey anti-mouse IgG and HRP-conjugated goat anti-mouse IgG (Fcγ-specific) were purchased from Jackson ImmunoResearch.

Flow cytometry and cell sorting. Flow cytometry analysis was done as previously described[7]. The samples were examined using a FACSCanto II (BD Biosciences) and the data were analysed using FlowJo software (Tree Star Software). CD4$^+$ T cells from iLNs were enriched using the CD4$^+$ T cell isolation kit (Miltenyi Biotech), the 8F10 T cells were sorted as CD45.2$^+$CD45.1$^-$CD4$^+$CD8$^-$B220$^-$CD11c$^-$ using FACSAria II (BD Biosciences).

CFSE and CMTMR labelling. For two-photon imaging, naive CD4 T cells were purified by two rounds of MACS negative selection using the naive CD4$^+$ T cell isolation kit (Miltenyi Biotech) to remove non-CD4 and CD44$^+$ T cells. The CD25$^+$ cells were further removed from the flow-through portion using the CD25 microbead kit (Miltenyi Biotech). The naive 8.3 or wild-type CD8 T cells were purified similarly by using the naive CD8$^+$ T cell isolation kit. Flow cytometry analysis confirmed that >95% of the cells were CD4$^+$/CD8$^+$CD25$^-$CD62L^{hi}CD44$^-$. CFSE (carboxyfluorescein succinimidyl ester) or 5-(and-6)-(((4-chloromethyl) benzoyl)amino)tetramethylrhodamine (CMTMR) labelling was performed using the Vybrant CFDA s.e. Cell Tracer Kit and the CellTracker Orange CMTMR Dye (both from ThermoFisher Scientific), respectively. In brief, T cells (10^7/ml in PBS) were incubated with 10 μM CFSE or 8 μM CMTMR for 25 min at 37 °C with a gentle shake after 10 min. Under these conditions, T cells were labelled with satisfactory intensities without significant cell death. Ice-cold PBS was then added to quench the labelling.

Adoptive transfer. CFSE- or CMTMR-labelled T cells were mixed 1:1 and were injected intravenously. For all the two-photon experiments, 2×10^6 T cells with either label were transferred. Varying T cell numbers in preliminary experiments determined that this amount resulted in a stable 0.5–0.8% reconstitution of the transferred T cells in the endogenous CD4 T cell pool, which was sufficient for two-photon imaging without causing obvious intraclonal competition. All the recipient mice were 3–4-week-old female mice unless otherwise mentioned. For experiments in Fig. 4f, FACS-sorted 8F10 T cells from pooled iLNs of 8–10 NOD or B16A mice were adoptively transferred intravenously into 4–6-week old NOD.*Rag1*$^{-/-}$ recipients (10^5 cells per mouse).

Two-photon imaging. Lymph nodes were removed, attached to coverslips, placed in CO_2-independent medium (Gibco) at room temperature and immediately imaged in a perfusion chamber to simulate blood flow (36.5 °C DMEM; 95% O_2 and 5% CO_2). Two-photon microscopy images were collected using a customized Leica SP8 Two-Photon Microscope (Leica Microsystems) equipped with a 25× and 0.95 numerical aperture water-immersion objective and a Mai Tai HP DeepSee Laser (Spectra-Physics) tuned to 840 nm. Fluorescence emission was guided directly to external hybrid photodetectors (Leica/Hamamatsu). For signal separation, we used three separate dichroic beam splitters without bandpass filters (Semrock): 484-nm edge BrightLine (FF484-FDi01), 495-nm edge BrightLine (FF495-Di03), and 560-nm edge BrightLine (FF560-Di01). The mirrors were arranged in dendritic fashion. Stacks were collected with 2.5 μm between images with 25–30 images per stack.

For cell tracking, two or three regions of one lymph node were randomly selected and cropped. Cells were tracked manually in 3D volume using Imaris 8.41 software (Bitplane). We tracked the first 11 time points from each track (representing 5 min and 10 velocities between the time points). Each dot represents the mean velocity out of the 10 that were tracked. We also calculated the meandering index and the motility coefficient for each track. Note that we chose tracks with the same length since the track length impacts these last two parameters. The meandering index and the motility coefficient data are not shown for space reasons however the results support the velocity data. The mean track velocities (μm/min) were calculated for individual tracks as previously described[11].

Surgical removal of pancreatic lymph nodes. NOD mice (3-week old) were anesthetized with a 4% mixture of isoflurane in oxygen. The two pLNs were exposed by gently retracting the spleen, pancreas, stomach and intestines, and were grasped with blunt forceps. Using an ophthalmic cautery on low power, the blood vessels on either side of the pLNs were cauterized and the pLNs were removed. The sham surgery was performed with the same procedures except that the pLNs were exposed without removal.

S961 administration. The S961 peptide (GSLDESFYDWFERQLGGGSGGSSLEEE WAQIQCEVWGRGCPSY) was synthesized by LifeTein, with an intrachain disulphide bridge between Cys33 and Cys40 (underlined). S961 (20 nMol/week) or control PBS was filled into the Alzet osmotic pump (2001 model, Durect) and inserted subcutaneously in the back of anaesthetized mice through an incision between scapula. Blood glucose levels were monitored twice a day (Chemstrip 2GP; Roche); mice with a level above 250 mg/dl for two consecutive measurements were considered diabetic.

Competitive ELISA assay. 96-well ELISA plates were coated with human insulin solution (1 μg/well) or peptides B:1–30 or B:9–23 (2 μM), and were blocked with 3% BSA overnight at 4 °C. Soluble competitive inhibitors, including different synthetic peptides and biological samples, were pre-incubated with the E11D7 (100 ng/ml), 6F3.B8 (20 ng/ml), or AIP (4 ng/ml) monoclonal antibodies for 30 min and the mixture was added to the plate-bound antigens for 1 h at room temperature. In the absence of soluble competitive inhibitors, these concentrations of the monoclonal antibodies resulted in about a 50% binding to the plate-bound antigens. HRP-conjugated goat anti-mouse IgG (1:10000) antibody was then added for 1 h; the responses were developed using the OptEIA TMB Substrate (BD). The data ($A_{450\,nm}$) were collected using an iMark Microplate Reader (Bio-Rad Laboratories). For quantifying the biological samples, each experiment was paired with a standard curve in which serially diluted amounts of soluble antigens were used to suppress the binding of their cognate monoclonal antibodies to the same antigen in the plate-bound form. The degree of inhibition by the biological samples was calculated relative to the blocking curve used by the specific antigen using an equation generated by linear regression of the standard curve.

Immunofluorescence microscopy. Mouse islets were isolated as previously described[7]. The islets were blocked with normal goat serum, fixed with 4% methanol-free formaldehyde, permeabilized with 0.2% saponin (Sigma), and stained with AIP or 6F3.B8 (50 μg/ml) for 45 min on ice. The samples were then stained with Alexa Fluor 594 F(ab)$_2$ donkey anti-mouse IgG (30 μg/ml), Alexa Fluor 647 Rabbit anti-insulin (20 μg/ml), and Alexa Fluor 488 anti-mouse CD11c (40 μg/ml) for 45 min on ice and mounted using the Prolong Diamond mountant (ThermoFisher). The samples were viewed using the Eclipse E800 microscope (Nikon) equipped with the EXi Blue fluorescence camera (Qimaging).

Electron microscopy with Immunogold. Islets were fixed in 4% paraformaldehyde (Polysciences) in 100 mM PIPES, 0.5 mM $MgCl_2$, pH 7.2 for 1 h at 4 °C. Samples were embedded in 10% gelatin and infiltrated overnight with 2.3 M sucrose/20% polyvinyl pyrrolidone in PIPES/$MgCl_2$ at 4 °C. Ultrathin sections of 50 nm were incubated with a blocking solution supplemented with 5% FBS and 5% normal goat serum for 30 min and subsequently incubated with rabbit anti-insulin (C27C9) and mouse anti-B chain (6F3.B8) antibodies for 1 h at room temperature. Sections were subsequently incubated with goat anti-mouse IgG conjugated to 18 nm colloidal gold and goat anti-rabbit IgG antibody conjugated to 12 nm colloidal gold for 1 h. Sections were stained with uranyl acetate and lead citrate and viewed on a JEOL 1200 EX transmission electron microscope (JEOL USA) equipped with an AMT 8-megapixel digital camera and AMT Image Capture Engine v.602 software (Advanced Microscopy Techniques). All labelling experiments were conducted in parallel with controls omitting the primary antibody. These controls were consistently negative.

Insulin secretion assay. Islets were equilibrated in DMEM supplemented with 10% FBS and 5.5 mM glucose for 24 h in 24-well plates. The medium was then replaced with 300 μl pre-warmed Krebs-Ringer-HEPES balance solution containing 0.2%

© 2018 Springer Nature Limited. All rights reserved.

BSA with 2.5 mM or 25 mM glucose. After 1 h incubation, the culture supernatants were collected for the competitive ELISA assay or mass spectrometry analysis.

β-cell granule isolation. Mouse and human islets were dispersed using non-enzymatic dispersion solution (Sigma). Cells were resuspended in PBS and lysed by passing them through a cell homogenizer (Isobiotec). The lysate was centrifuged twice for 10 min at 500*g*, 4 °C to pellet cell debris. The supernatant was centrifuged for 10 min at 5,000*g*, 4 °C. The 5,000*g* spin was repeated on the supernatant and the two pellets were combined. This fraction was highly enriched in peptide-containing vesicles compatible with the crinophagic bodies, and as such have been labelled. This fraction may also contain organelles other than the insulin-containing ones. The supernatant after the 5,000*g* spin was centrifuged for 30 min at 25,000*g*, 4 °C to pellet secretory granules. This supernatant was discarded, and the 25000*g* pellet was suspended in 100 μl PBS. The microcentrifuge used for granule isolation was an Eppendorf 5417R (Eppendorf) with a FA45-24-11 fixed angle rotor. Fractions were frozen at −80 °C and thawed at 37 °C for five cycles to release the contents of granules. After freeze-thaw, complete protease inhibitor cocktail was added to the sample which was then concentrated by SpeedVac to <100μl. The sample was passed through C18 Ziptips (Pierce) and peptides then were eluted in 0.1% formic acid/95% acetonitrile and then dried with a SpeedVac.

Sample preparation for mass spectrometry analysis. Biological samples were treated with 2.5% trifluoroacetic acid (TFA) to a final concentration of 0.36% (v/v), and the peptides were purified using the C18 Ziptips, eluted with 0.1% formic acid in 95% acetonitrile, and lyophilized. For peptide capture, TFA-adjusted mouse urine (12 ml) was cleaned up using C18 Sep Pak cartridges (Waters). The analytes retained by the cartridge sorbent were eluted with methanol, lyophilized, and reconstituted with 2 ml sterile PBS. The material was then incubated with a 1:1 mixture of sepharose pre-conjugated with AIP or 6F3.B8 monoclonal antibodies (1 ml slurry total) for 72 h at 4 °C with gentle rotation. The urine–sepharose mixture was poured into a Bio-Rad Econo column, and after extensive washing, the antibody-bound material was eluted with 10% acetic acid and lyophilized.

Mass spectrometry. A Dionex UltiMate 1000 system (Thermo Scientific) was coupled to an Orbitrap Fusion Lumos (Thermo Scientific) through an EASY-Spray ion source (Thermo Scientific). Peptide samples were loaded (30 μl/min, 1 min) onto a trap column (100 μm × 2 cm, 5 μm Acclaim PepMap 100 C18, 50 °C), eluted (300 μl/min) onto an EASY-Spray PepMap RSLC C18 column (2 μm, 25 cm × 75 μm ID, 50 °C, Thermo Scientific) and separated with the following gradient, all % Buffer B (0.1% formic acid in ACN): 0–40 min, 2–22%; 40–50 min, 22–35%; 50–60 min, 35–95%; 60–70 min, isocratic at 95%; 70–71 min, 95–2%, 71–85 min, isocratic at 2%. Spray voltage was 1900 V, ion transfer tube temperature was 275 °C, and RF lens was 30%. Mass spectrometry scans were acquired in profile mode and MS/MS scans in centroid mode, for ions with charge states 2–7, with a cycle time of 3 s. For HCD, mass spectra were recorded from 375–1500 Da at 120K resolution (at $m/z = 200$), and MS/MS was triggered above a threshold of 2.5×10^4, with quadrupole isolation (1.6 Da) at 30K resolution, and collision energy of 30%. Dynamic exclusion was used (35 s). For high SA EThcD, mass spectra were acquired from 350–1500 Da at 60K resolution, and MS/MS spectra were triggered for ions above a threshold of 5×10^4 with quadrupole isolation (0.7 Da) at 15K resolution. Fragmentation employed calibrated charge-dependent ETD, with SA (40%) applied in the HCD cell. Dynamic exclusion was used (60 s). For low SA EThcD, mass spectra were recorded from 375–1500 Da at 120K resolution (at $m/z = 200$), and MS/MS spectra were acquired for ions above a minimum intensity threshold of 2.5×10^4 at 15K resolution. ETD reaction time was fixed at 100 ms, with SA (15%) applied in the HCD collision cell.

Mass spectrometry data analysis. Data files were uploaded to PEAKS 8.0 (Bioinformatics Solutions) for processing, de novo sequencing and database searching. Resulting sequences were searched against the UniProt Mouse Proteome database (downloaded 8 June 2017; 25,144 entries) with mass error tolerances of 20 ppm and 0.02 Da for parent and fragment, respectively, no enzyme specificity and no fixed or variable modifications. The Common Repository for Adventitious Proteins database (www.thegpm.org/crap/) was used to identify contaminant proteins. FDR estimation was enabled. Peptides were filtered for $-10\log P \geq 20$, and proteins were filtered for $-10\log P \geq 30$ and one unique peptide. For all experiments, this gave an FDR of <1% at the peptide-spectrum match level. Peptides matching to insulin-1 and insulin-2 were manually verified by visual inspection. For relative quantification, peak areas for all manually verified peptides were exported from PEAKS, normalized to the total ion current, and $\log_2$ transformed.

T cell stimulation and antigen presentation assay. In Extended Data Fig. 1i, ConA-stimulated peritoneal macrophages were treated with 0.2 or 1 μM S961 for 1 h at 37 °C and were then cultured with the IIT-3 T cell hybridoma that recognizes the 13–21 peptide, in the presence of serially diluted insulin (I9278; Sigma). In Extended Data Fig. 2e, C3g7 cells were treated with chloroquine for 2 h at 37 °C, washed, pulsed with the antigens, and cultured with T cell hybridomas. After incubation for 18 h, the culture supernatants were assayed for IL-2 production.

Bone marrow chimaera. The female donor 8F10.TCRα$^{-/-}$ (CD45.2) mice were injected intraperitoneally with fluorouracil (200 mg/kg), and bone marrow cells were isolated from the femur and tibia on day 5. The cells were adoptively transferred into sublethally irradiated (600 rads) 3-week old female NOD or B16A hosts (10^4/mouse).

RNA-seq analysis. Total RNA was isolated using the Ambion RNAqueous-Micro kit (Thermo Fisher Scientific). RNA-seq library preparation and sequencing was performed as previously described[32]. The differential expression analysis was done with the DESeq2 package (version 1.18.1). Multifactor analysis was used to account for donor effect. Specifically, paired 8F10-NOD and 8F10-B16A samples from one isolation (four pairs in total) were treated as one donor group. Gene set enrichment pathways analysis was done using the Broad Institute's GSEA software and MSigDB Hallmark or C7 immunological signatures databases. The latter included datasets: GSE28726[24], GSE1000001_1577_200_UP[25], GSE9650[26] and GSE32025[28]. All heat maps are in $\log_2$ scale. The gene expression matrix counts were adjusted for donor effect with Combat (sva package) only for heat maps and clustering.

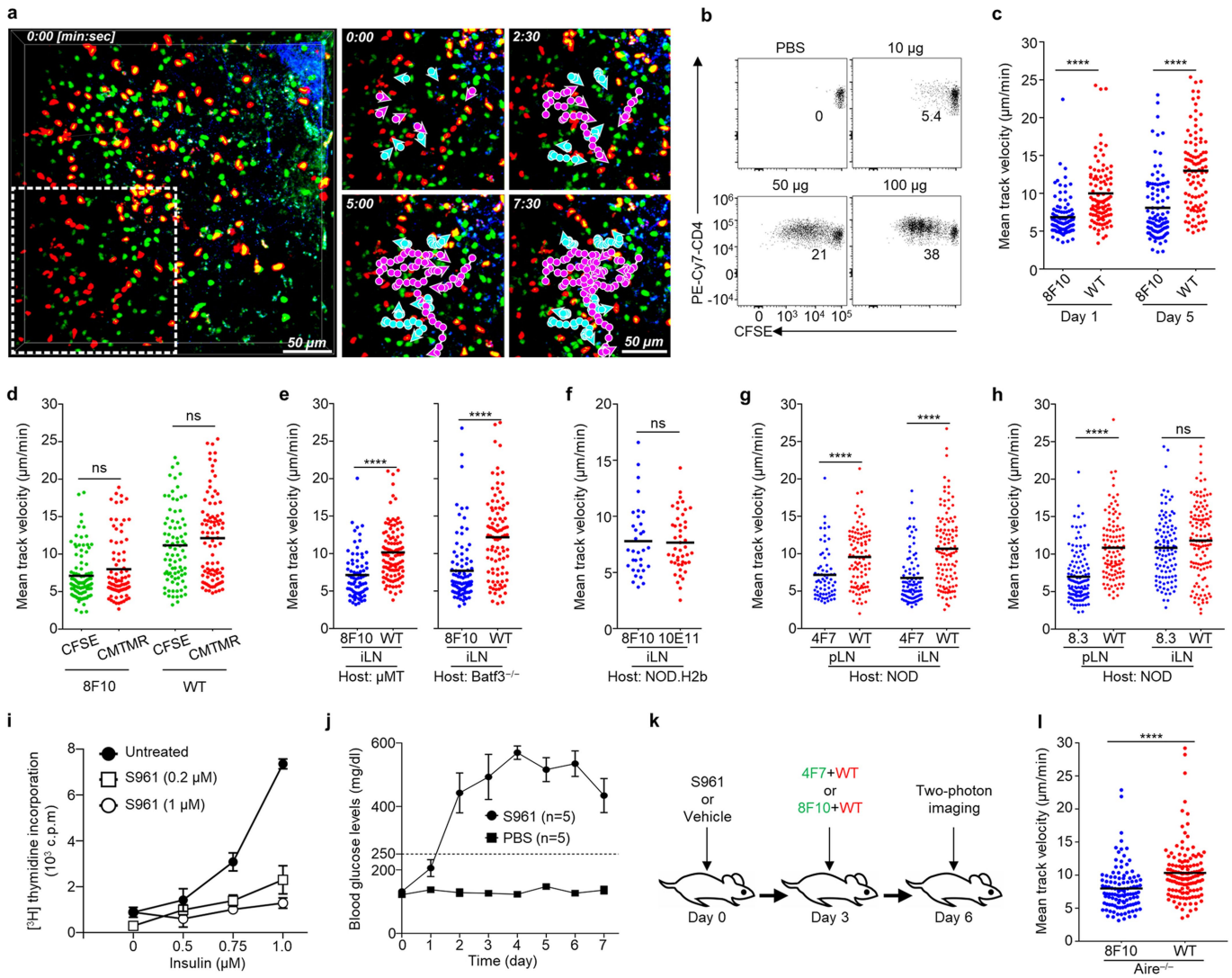
Statistics. Mice were age and gender matched. Among the mice with matched ages and genders, they were randomized and distributed equally into experimental groups. Power analysis was used to estimate the sample size in some experiments, as described in the Reporting Summary. The investigators were not blinded to allocation during experiments and outcome assessment. One-way ANOVA with Sidak's multiple comparisons test was used to determine significant differences among multiple groups with unpaired biological replicates. The two-tailed unpaired Student's *t*-test was used to determine significant differences between two groups with unpaired biological replicates. The two-tailed paired Student's *t*-test was used to calculate *P* values of each pair of independent experiments. The log-rank test was used to determine the significant difference of diabetes incidence.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The RNA-seq data have been deposited in the Gene Expression Omnibus under accession number GSE114824. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD009919.

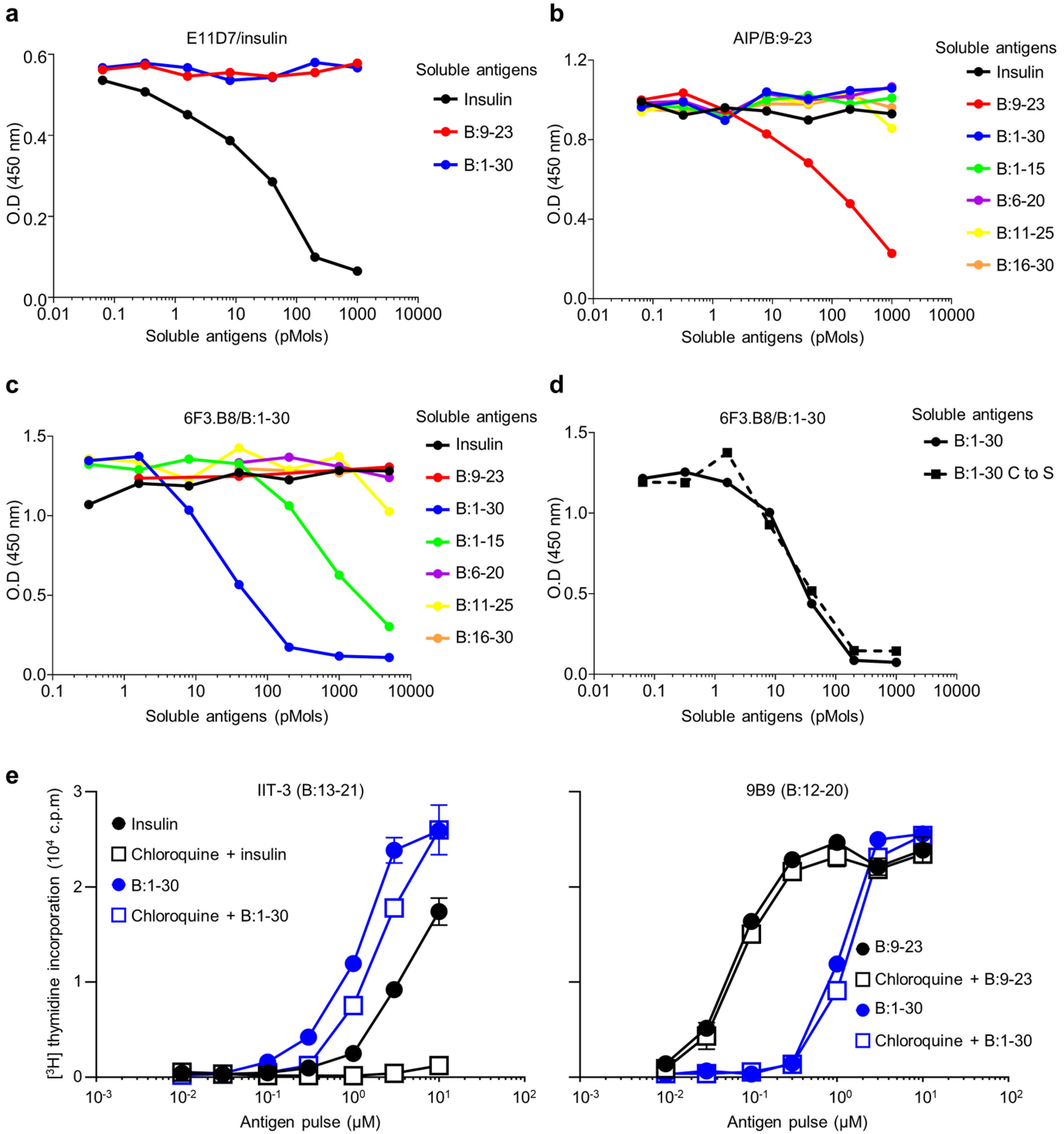
31. Aly, H. et al. A novel strategy to increase the proliferative potential of adult human β-cells while maintaining their differentiated phenotype. *PLoS One* **8**, e66131 (2013).
32. Ferris, S. T. et al. The islet-resident macrophage is in an inflammatory state and senses microbial products in blood. *J. Exp. Med*. **214**, 2369–2385 (2017).

© 2018 Springer Nature Limited. All rights reserved.



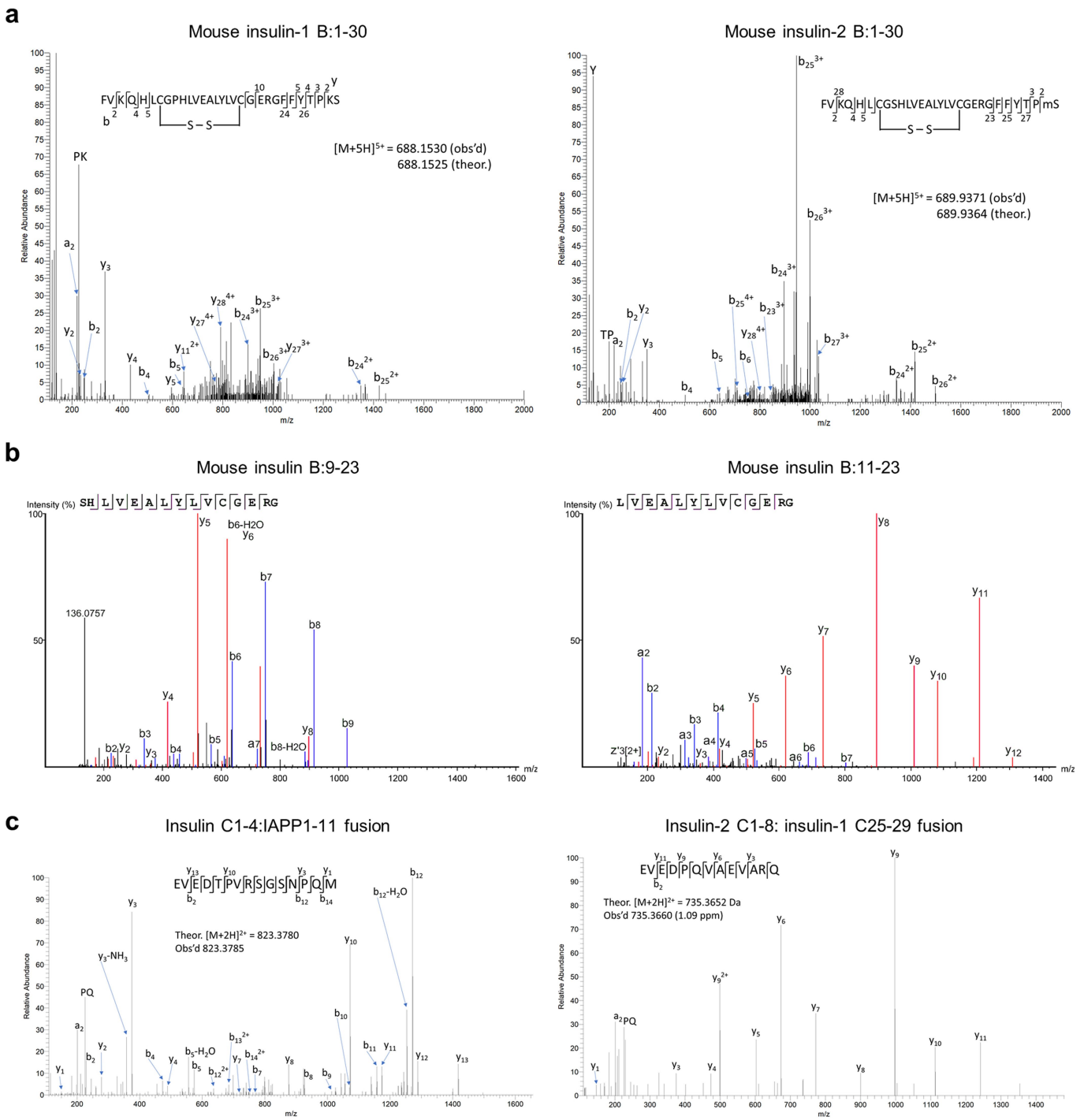
Extended Data Fig. 1 | Probing peripheral antigen presentation by two-photon imaging; the motility assay. a, Representative 3D reconstructions of two-photon *z*-stacks visualizing CFSE-labelled anti-HEL 10E11 TCR transgenic and CMTMR-labelled wild-type CD4 T cells in an iLN explant on day 3 post transfer. Individual T cells were tracked in the area bound by the dashed line. Right, magnified views of this region, showing movement of T cells over a 7.5-min time interval. Quantification was performed over a 5-min interval. Cyan and purple tracks represent 10E11 and wild-type T cells, respectively. Mice were injected with 10 µg HEL. **b**, NOD mice (CD45.1) were injected intraperitoneally with indicated amounts of HEL. Six hours after injection, naive CFSE-labelled 10E11 (CD45.2) T cells were transferred. On day 3, CFSE dilution of the transferred T cells ($CD45.2^+CD45.1^-CD4^+V\beta8.1/8.2^+$) in the iLNs was measured by flow cytometry. Data are representative of two independent experiments. **c**, Mean track velocities of 8F10 and wild-type CD4 T cells in iLNs from NOD recipients on day 1 or day 5 post transfer. **d**, CFSE(8F10) plus CMTMR(WT) or CMTMR(8F10) plus CFSE(WT) T cells were separately transferred into two cohorts of NOD recipients, and their mean track velocities in iLNs on day 3 were compared by paired two-photon imaging analysis. **e**, Mean track velocities of 8F10 and wild-type CD4 T cells in NOD.µMT or NOD.$Batf3^{-/-}$ recipients on day 3 post transfer. **f**, Mean track velocities of 8F10 and 10E11 T cells in NOD.H2b recipients 24 h post transfer. **g**, **h**, Mean track velocities of 4F7 and wild-type CD4 (**g**) or 8.3 and wild-type CD8 (**h**) T cells in NOD recipients on day 3 post transfer. **i**, Response (mean ± s.e.m.) of the B:13–21-specific IIT-3 T cells to ConA-activated peritoneal macrophages treated with or without S961 before insulin pulse. **j**, Blood glucose levels (mean ± s.e.m.) of 3-week old NOD mice infused with S961 or PBS via osmotic pumps. **k**, The scheme of the experiments in Fig. 1h, i. **l**, Mean track velocities of 8F10 and wild-type CD4 T cells in iLNs of $Aire^{-/-}$ recipients. Data summarize two (**c**, **d**, **f**, **l**) or three (**e**, **g**, **h**) independent experiments. Each dot represents individual T cell tracks, and the bar denotes the mean. ns, not significant; $^{****}P < 0.0001$; one-way ANOVA with Sidak's multiple comparisons test (**c**, **d**, **g**, **h**) or two-tailed unpaired Student's *t*-test (**e**, **f**, **l**).

© 2018 Springer Nature Limited. All rights reserved.



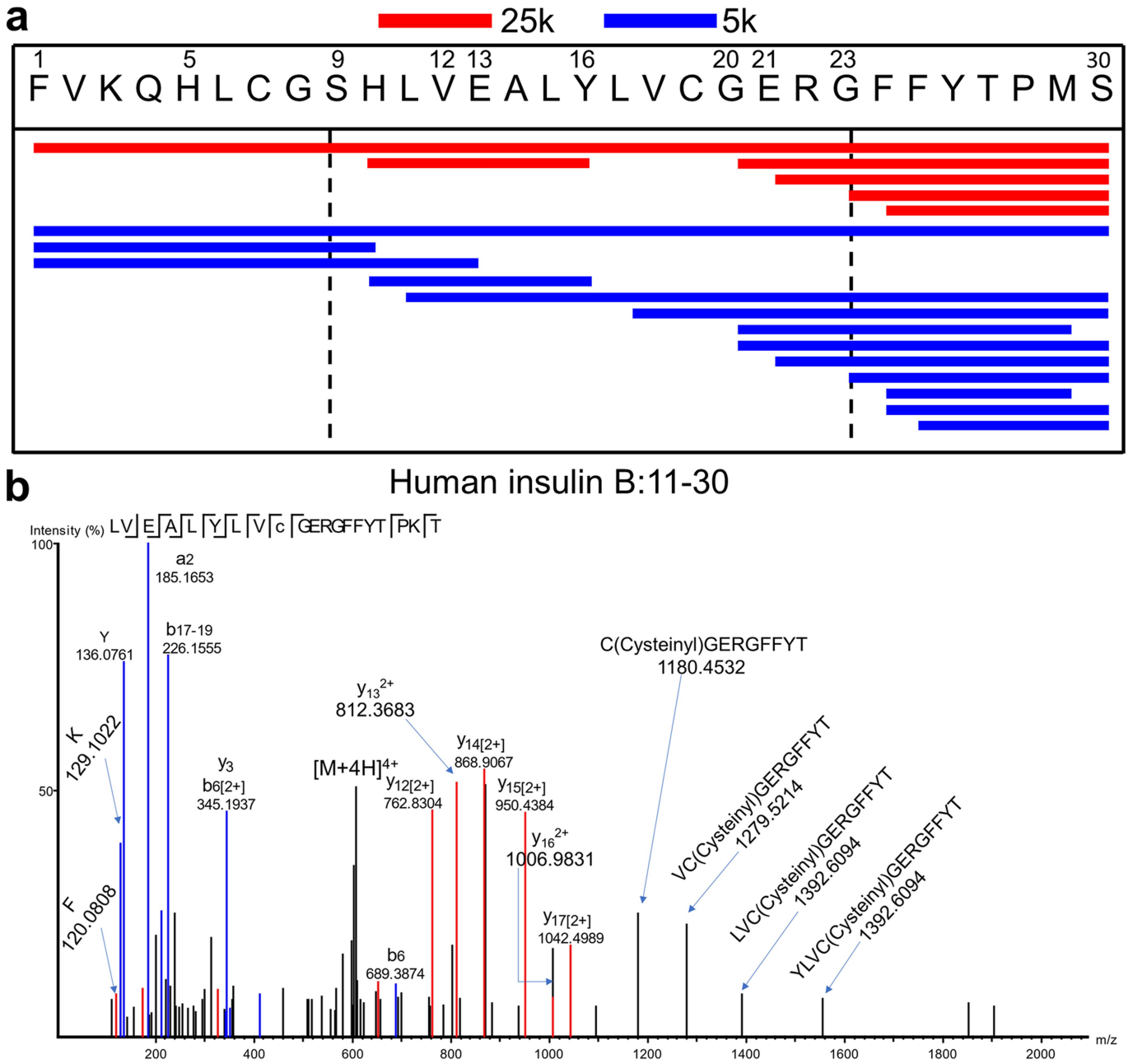
Extended Data Fig. 2 | Analysis of insulin peptide-specific monoclonal antibodies and presentation of the intact B-chain. a–**c**, Competitive ELISA responses showing the binding of: anti-insulin monoclonal antibody (E11D7) to plate-bound insulin (**a**) anti-B:9–23 monoclonal antibody (AIP) to plate-bound B:9–23 (**b**), and anti-B:1–30 monoclonal antibody (6F3.B8) to plate-bound B:1–30 (**c**) in the presence of serial dilutions of the indicated soluble antigens as a competitive inhibitor. Inhibition by a specific soluble antigen indicates the specificity of the monoclonal antibody to this antigen. **d**, Competitive ELISA responses showing the binding of 6F3.B8 to plate-bound B:1–30 in the presence of soluble unmodified B:1–30 or B:1–30 in which the two cysteines were changed to serines (B:1–30 C to S). The results indicate the intrachain link formed by the cysteines does not influence the specificity of the 6F3. B8 monoclonal antibody. Data are means representing two independent experiments. **e**, Responses of the B:13–21-specific IIT-3 (left) or the B:12–20-specific 9B9 (right) T cell hybridoma to C3g7 APCs treated with or without 100 μM chloroquine for 2 h and pulsed with indicated antigens after extensive washes. C3g7 cells are a B cell lymphoma line expressing I-A^{g7}, and are used as APCs. The results of the effects of chloroquine indicate that reactivity to insulin, but not to B:9–23 or B:1–30 require internal processing. Data are mean ± s.e.m., representative of two independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



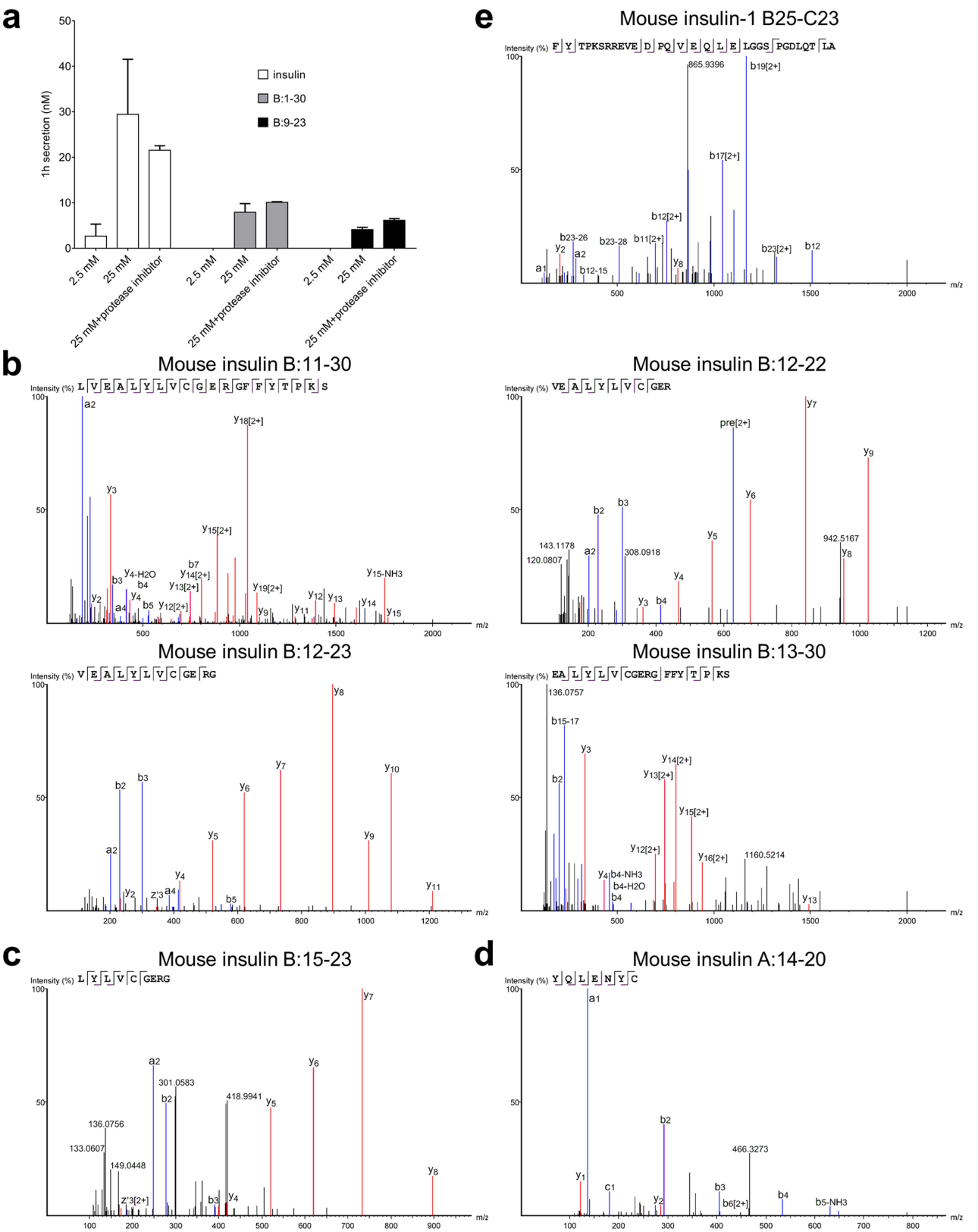
Extended Data Fig. 3 | nLC–MS/MS analysis of mouse β-cell granules. **a**, Mass spectra of mouse insulin-1 B:1–30 with intramolecular disulphide bonds (left) and mouse insulin-2 B:1–30 with oxidized methionine in position 29 (right). **b**, Mass spectra of mouse insulin B:9–23 (left) and B:11–23 (right), which were exclusively identified in the 5k granules of B6g7, B6 and NOD mice. **c**, Mass spectra of two hybrid peptides identified in the 5k granules. The sequence (EVEDTPVRSGSNPQM, left) represents a C-peptide (underlined)–islet amyloid polypeptide (IAPP) fusion, and the sequence (EVEDPQVAEVARQ, right) represents a fusion of the N terminus of insulin-2 C-peptide (underlined) with the C terminus of insulin-1 C-peptide.

© 2018 Springer Nature Limited. All rights reserved.



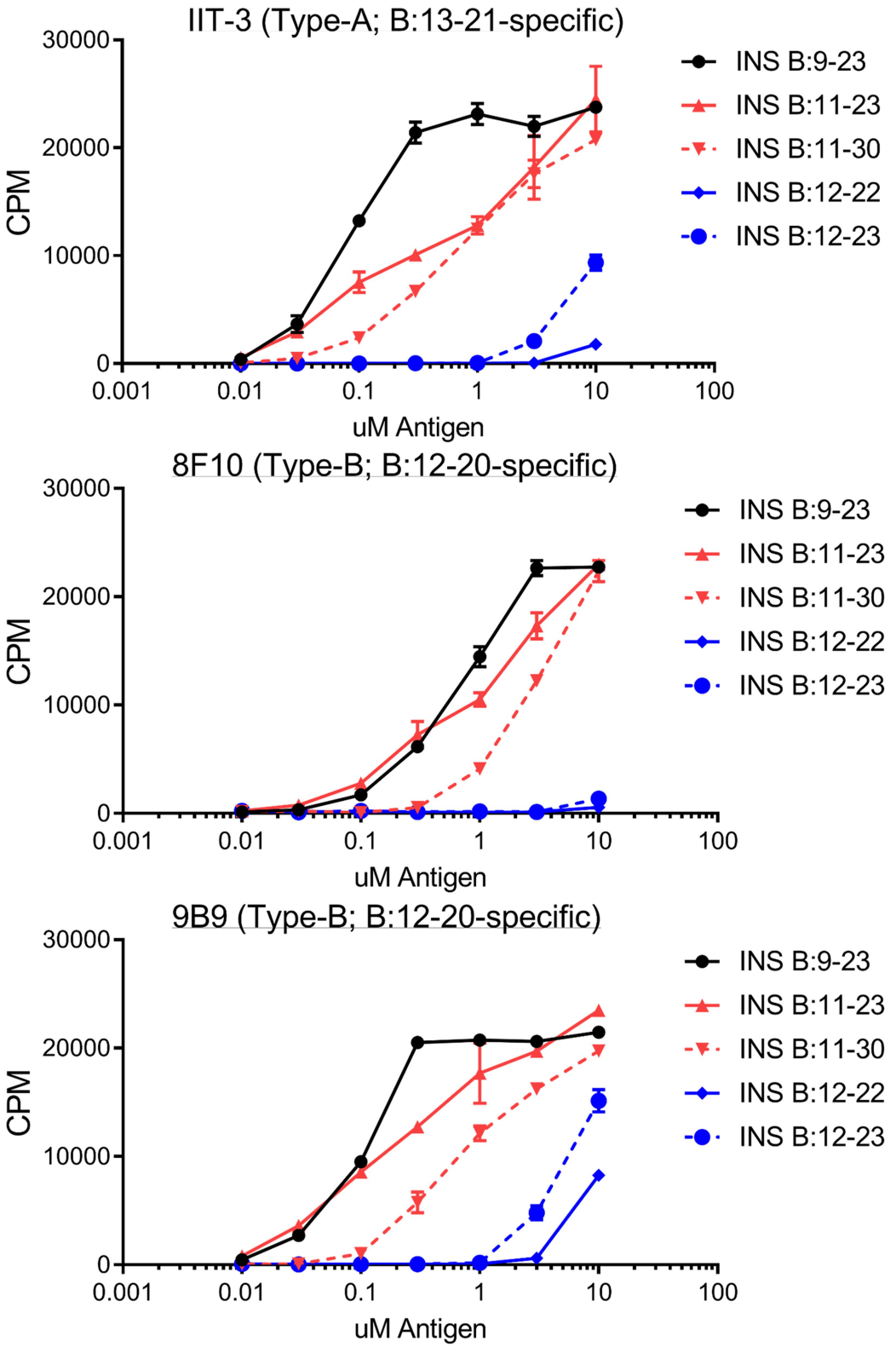
Extended Data Fig. 4 | nLC–MS/MS analysis of human β-cell granules. **a**, Peptide coverage of insulin B chain identified in human 25k (red) and 5k (blue) β-cell granules using nLC–MS/MS analysis. Shown is the alignment of individual peptides (each line) with the human insulin B:1–30 segment. Data summarizes results from four independent runs using human islets from three individual donors. **b**, A mass spectrum showing a sequence representing human insulin B:11–30 that was identified in the 5k granules. The cysteinylation in position 19 is indicated.

© 2018 Springer Nature Limited. All rights reserved.



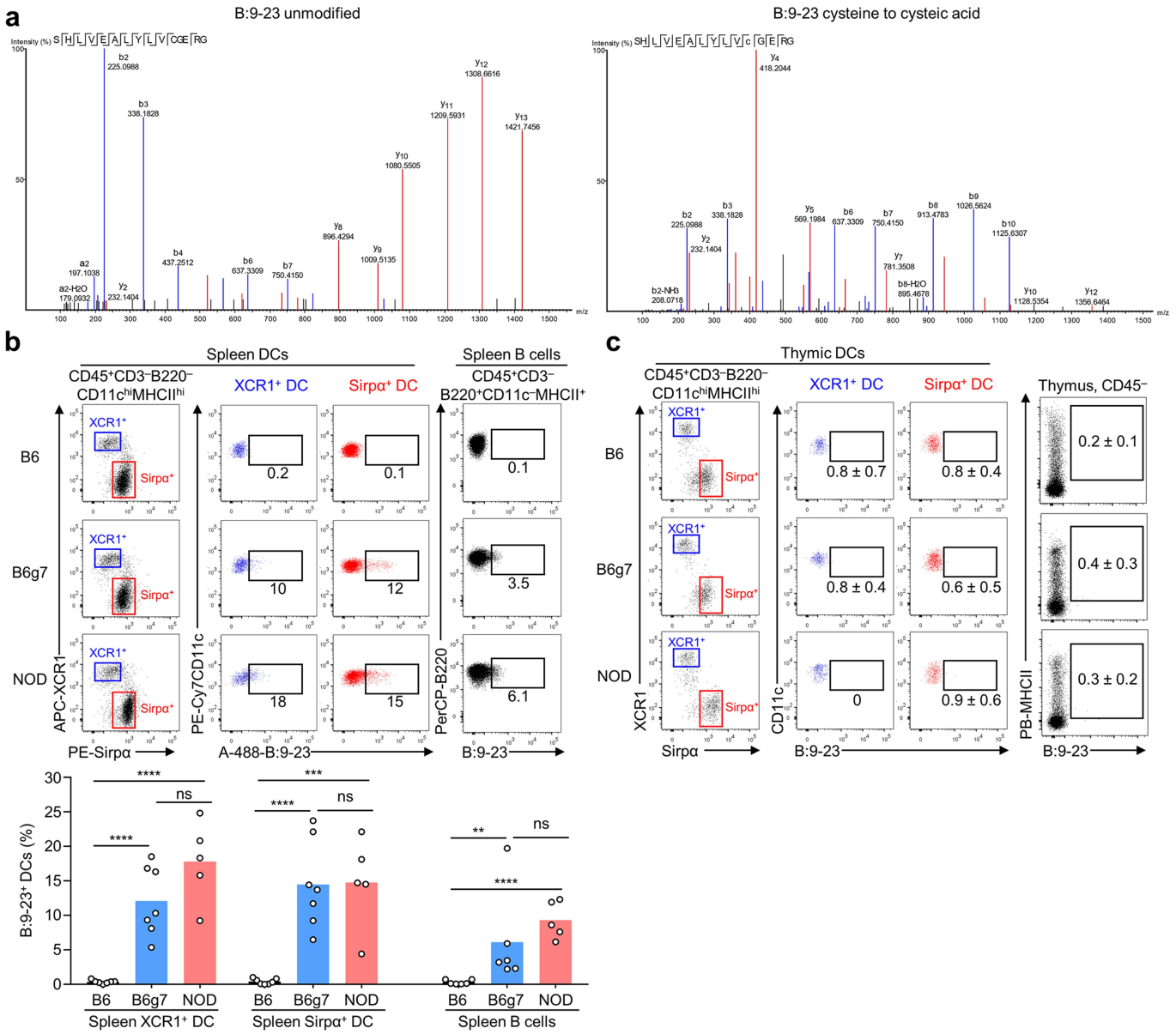
Extended Data Fig. 5 | Analysis of insulin peptides secreted from islets upon glucose challenge. a, Insulin secretion assay was performed as described in Fig. 3a–c, except that protease inhibitors were added during the 25-mM glucose challenge. The supernatants were then collected for the competitive ELISA assay. Data are mean ± s.e.m. from two independent experiments. **b**, Mass spectra of four secreted peptides that contain the B:12–20 and/or B:13–21 epitopes as listed in Fig. 3e. Secreted B:1–30 sequences are identical to those in Extended Data Fig. 3a, and B:9–23 and B:11–23 share identical sequences with those in Extended Data Fig. 3b. **c**, A mass spectrum of the secreted insulin B:15–23 MHC-I (K^d)-binding peptide. **d**, A mass spectrum of the secreted insulin A:14–20 MHC-I (D^b)-binding peptide. **e**, A mass spectrum showing a representative B–C-spanning peptide (B25–C23).

© 2018 Springer Nature Limited. All rights reserved.



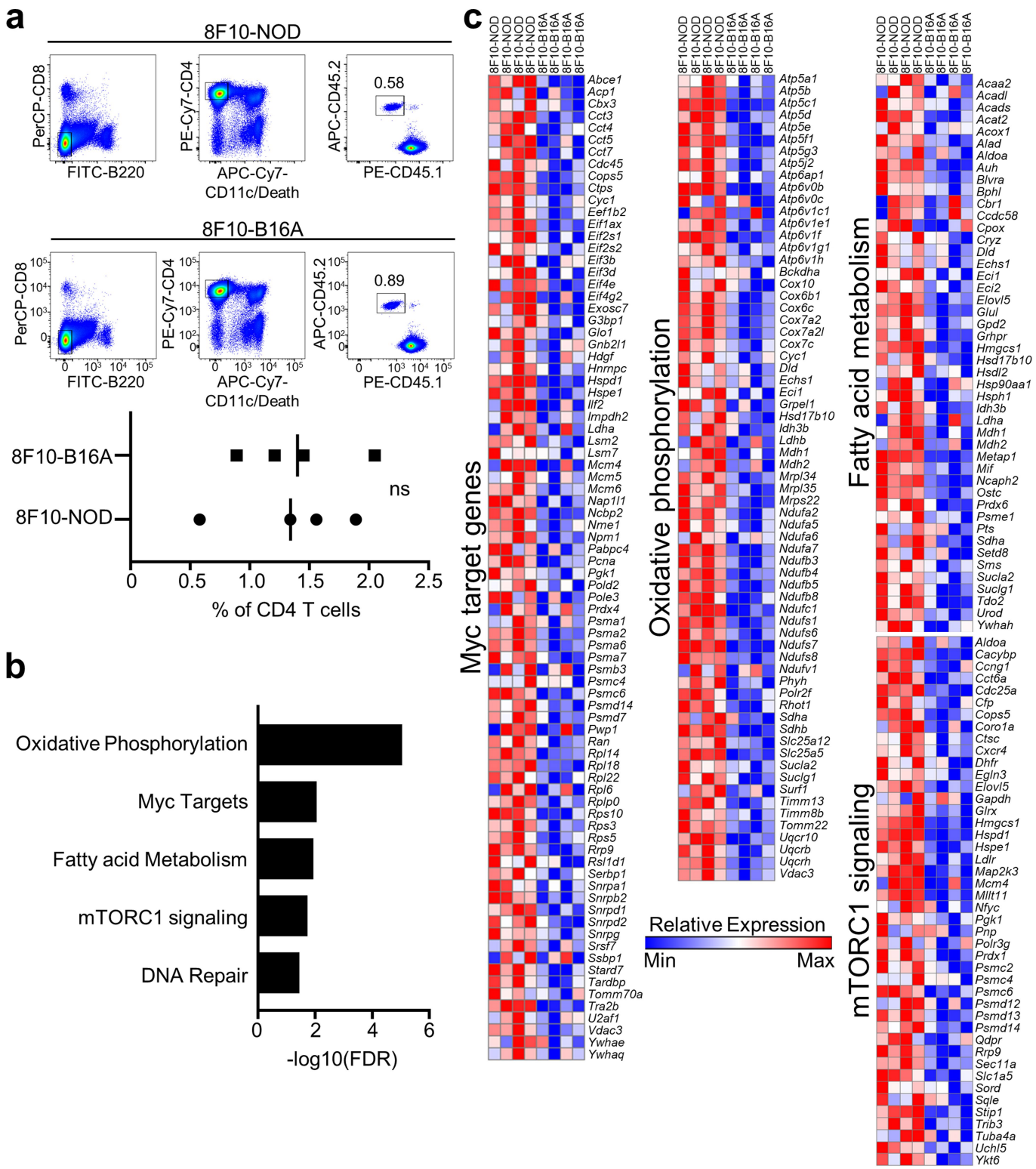
Extended Data Fig. 6 | T cell responses to B:9–23-associated peptides. Responses of three insulin-reactive T cell hybridomas to insulin peptides associated with the 9–23 region of the B chain as identified in Fig. 3e. The C3g7 cells were used as APCs. Data are mean ± s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



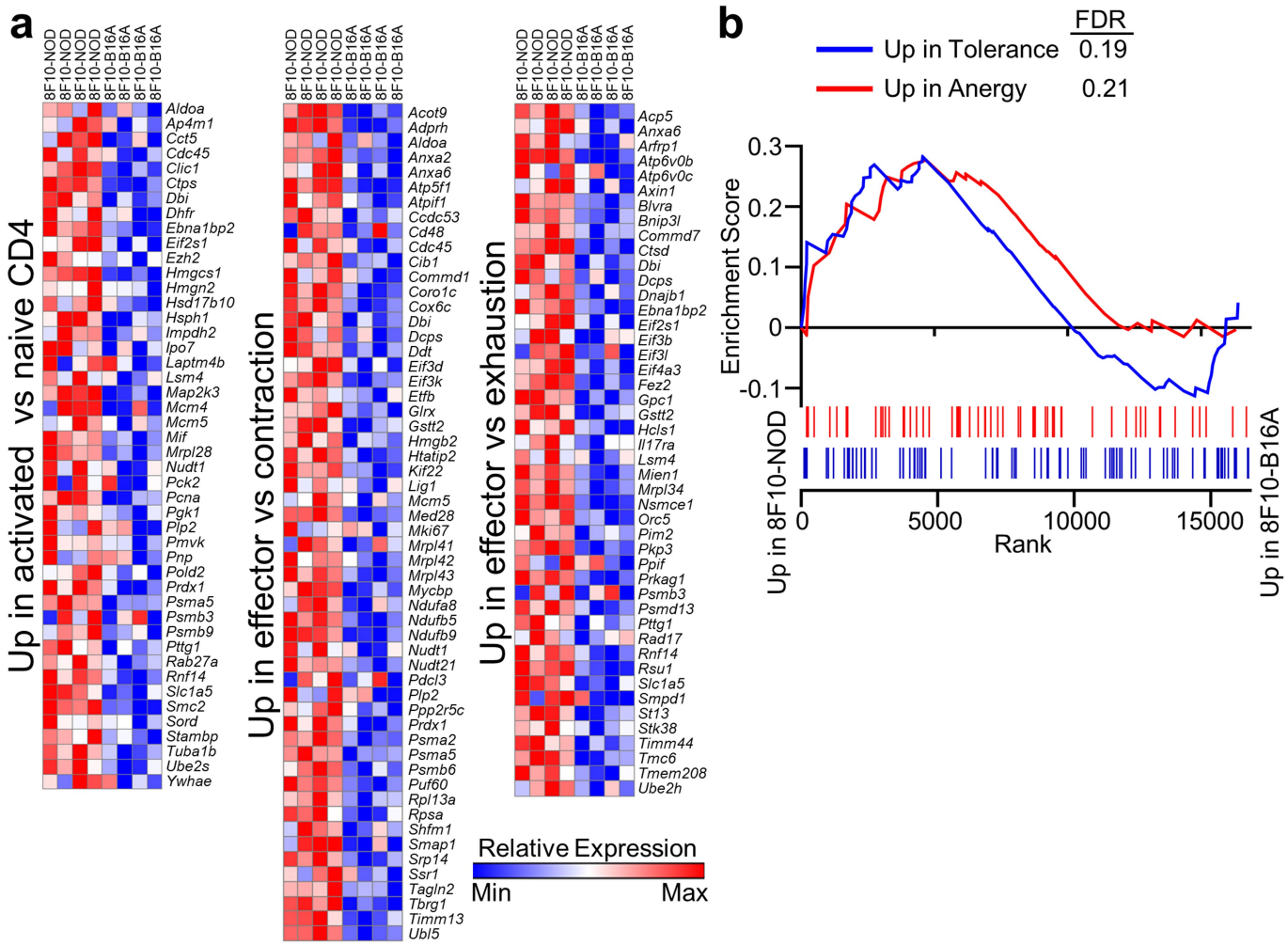
Extended Data Fig. 7 | Characterization of circulating B:9–23 and its localization into lymphoid organs. a, Unmodified synthetic B:9–23 (3 pmol) was spiked into 1 ml PBS, purified using C18 tips, lyophilized, and analysed by nLC–MS/MS. The data show the appearance of unmodified B:9–23 (left) together with oxidation of Cys19 to cysteic acid (right). **b**, **c**, Alexa Fluor 488-conjugated B:9–23 peptide (100 μg) was injected intravenously into 4-week old B6, B6g7 and NOD mice. An hour later, spleens and thymi were harvested, digested with liberase and DNase, and binding to splenic and thymic APCs was measured by flow cytometry. **b**, Representative FACS plots showing the binding of B:9–23 to splenic XCR1$^+$ and Sirpα$^+$ dendritic cell (DC) subsets and B cells (top). The bar graph summarizes cumulative results from individual mice (each point), pooled from three independent experiments. ns, not significant; $^{**}P < 0.05$; $^{***}P < 0.01$; $^{****}P < 0.005$, two-tailed unpaired Student's *t*-test. **c**, Representative FACS plots showing the binding of B:9–23 to thymic XCR1$^+$ and Sirpα$^+$ DC subsets and to CD45$^-$ cells expressing MHCII. Data are mean ± S.D from five individual mice per strain from two independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



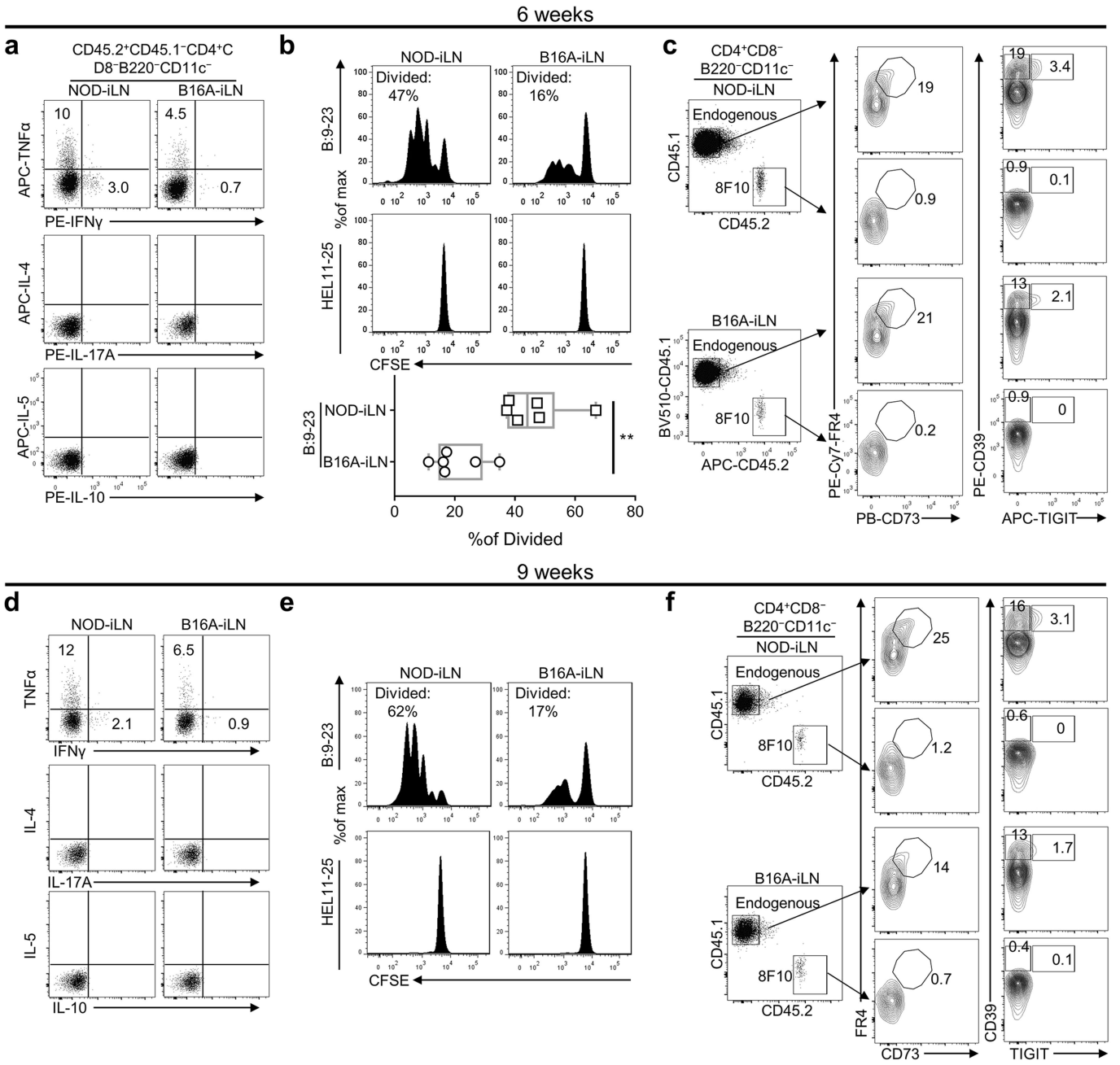
Extended Data Fig. 8 | RNA-seq analysis of 8F10 T cells developed in NOD or B16A hosts. a, Representative FACS plots (top) showing the sorting strategy and recovery of 8F10 T cells from iLNs of NOD or B16A-recipient mice six weeks after adoptive transfer of bone marrow. The scatter plot (bottom) shows the percentage of recovered 8F10 T cells among total CD4 T cells from four independent experiments. ns, not significant; two-tailed paired Student's *t*-test. **b**, Biological pathways that are significantly enriched in the 8F10-NOD versus 8F10-B16A samples using GSEA and Hallmark database. **c**, Heat maps of all enriched genes in individual metabolic pathways depicted in Fig. 4c.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | 8F10 T cells exhibit an effector phenotype, but no anergy or exhaustion phenotype, at the transcription level during peripheral antigen recognition. a, Heat maps showing all the enriched genes of the three immunological pathways illustrated in Fig. 4d. **b**, GSEA enrichment plots performed on differentially expressed genes in 8F10 T cells from the NOD-iLN versus B16A-iLN condition using datasets characterizing CD4 T cell anergy and CD8 T cell tolerance.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Functional analysis of 8F10 T cells developed in NOD or B16A hosts. a–f, The bone marrow chimaera was constructed as in Fig. 4a, and T cells were examined after 6 (**a**–**c**) or 9 (**d**–**f**) weeks. **a**, **b**, **d**, **e**, Bulk $CD4^+$ T cells were purified from iLNs of individual NOD or B16A mice (three per group) by two rounds of MACS negative selection. To examine cytokine repertoire (**a**, **d**), half of the individual T cell samples were combined. The remainder were kept as individual samples, labelled with CFSE (1.5 μM), and used to measure cell proliferation (**b**, **e**). In either case, T cells were mixed with NOD.*Rag1*$^{-/-}$ splenocytes (1:2 ratio) and stimulated with B:9–23 for 16 (**a**, **d**) or 72 (**b**, **e**) hours. **a**, Representative FACS plots showing intracellular cytokine staining of the 8F10 T cells from NOD-iLN or B16A-iLN, after stimulation with B:9–23 for 16 h (brefeldin A was added for the last 4 h). Production of IL-4, IL-17A, IL-5 and IL-10 was not detected. Data are representative of two independent experiments with 3 mice combined per experiment. **b**, Representative FACS plots (top) showing CFSE dilution of the 8F10 T cells stimulated by B:9–23 or the control HEL11–25 peptide for 72 h. The results of 6 individual mice from two independent experiments are summarized in the box plots (bottom). Box plots show the median, box edges represent the first and third quartiles, and the whiskers extend from the minimum to the maximum. $^{**}P < 0.01$, two-tailed unpaired Student's *t*-test. **c**, Representative FACS plots showing ex vivo surface staining of FR4 and CD73 as well as CD39 and TIGIT on endogenous $CD4^+$ or 8F10 T cells in the iLNs of NOD or B16A mice. Data are representative of three mice analysed in two independent experiments. **d**–**f**, Experiments were performed in week 9 following the procedures described in **a**–**c**. The data in **d**–**f** are from a single experiment.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0324-7

53BP1–RIF1–shieldin counteracts DSB resection through CST- and Polα-dependent fill-in

Zachary Mirman[1,5], Francisca Lottersberger[1,4,5], Hiroyuki Takai[1], Tatsuya Kibe[1], Yi Gong[1], Kaori Takai[1], Alessandro Bianchi[1,2], Michal Zimmermann[1,3], Daniel Durocher[3] & Titia de Lange[1]*

In DNA repair, the resection of double-strand breaks dictates the choice between homology-directed repair—which requires a 3′ overhang—and classical non-homologous end joining, which can join unresected ends[1,2]. BRCA1-mutant cancers show minimal resection of double-strand breaks, which renders them deficient in homology-directed repair and sensitive to inhibitors of poly(ADP-ribose) polymerase 1 (PARP1)[3–8]. When BRCA1 is absent, the resection of double-strand breaks is thought to be prevented by 53BP1, RIF1 and the REV7–SHLD1–SHLD2–SHLD3 (shieldin) complex, and loss of these factors diminishes sensitivity to PARP1 inhibitors[4,6–9]. Here we address the mechanism by which 53BP1–RIF1–shieldin regulates the generation of recombinogenic 3′ overhangs. We report that CTC1–STN1–TEN1 (CST)[10], a complex similar to replication protein A that functions as an accessory factor of polymerase-α (Polα)–primase[11], is a downstream effector in the 53BP1 pathway. CST interacts with shieldin and localizes with Polα to sites of DNA damage in a 53BP1- and shieldin-dependent manner. As with loss of 53BP1, RIF1 or shieldin, the depletion of CST leads to increased resection. In BRCA1-deficient cells, CST blocks RAD51 loading and promotes the efficacy of PARP1 inhibitors. In addition, Polα inhibition diminishes the effect of PARP1 inhibitors. These data suggest that CST–Polα-mediated fill-in helps to control the repair of double-strand breaks by 53BP1, RIF1 and shieldin.

This study was initiated to determine whether the control of 5′ resection at double-strand breaks (DSBs) resembles the regulation of resection at telomeres. The formation of telomeric t-loops requires the generation of 3′ overhangs after DNA replication[12–15]. Newly replicated telomeres are resected by EXO1, which generates 3′ overhangs that are too long and require fill-in mediated by Polα–primase[16] (Fig. 1a). Polα–primase is brought to telomeres by an interaction between CST (also known as α-accessory factor, AAF[11,17]) and POT1B in mouse shelterin[16] (Fig. 1a). Here we test whether CST–Polα fill-in of 3′ overhangs has a role in the regulation of DSB resection by 53BP1, RIF1 and shieldin.

To study the role of CST at sites of DNA damage we used telomeres that lack shelterin protection, which are a model system for DSB resection[8,15,18–20]. Hyper-resection occurs upon Cre-mediated removal of TPP1 (and POT1A and POT1B) from telomeres of *Tpp1*$^{f/f}$ mouse embryo fibroblasts (MEFs). This hyper-resection is counteracted by 53BP1 and RIF1[20], which accumulate in response to ATR signalling at telomeres that lack POT1A (Fig. 1a). As does 53BP1, shieldin limited hyper-resection at telomeres that lack TPP1: TPP1-deficient cells that lacked either REV7 or SHLD2 showed telomere hyper-resection (Fig. 1b–d, Extended Data Fig. 1a–c).

As CST is essential[21,22], we used short hairpin RNAs (shRNAs) to explore the role of CST in telomere hyper-resection. Depletion of STN1 or CTC1 increased the telomeric overhang signal in cells that lack TPP1 (Fig. 1b–d, Extended Data Figs. 1d, 2), and tests with *Escherichia coli* ExoI confirmed that the signal derived from a 3′ overhang (Extended Data Fig. 1e, f). Knockdown of STN1 or CTC1 did not affect the resection at telomeres when TPP1 was deleted from REV7-deficient cells (Fig. 1b–d, Extended Data Fig. 2). Furthermore, STN1 knockdown had no effect on telomere hyper-resection when either 53BP1 or RIF1 were absent, or when cells contained a form of 53BP1 that does not recruit RIF1[23] (Extended Data Fig. 3). These data suggest that CST acts in a 53BP1-, RIF1- and shieldin-dependent manner to limit the formation of single-stranded DNA at dysfunctional telomeres.

To determine whether CST also counteracted resection at sites of ATM signalling, we used conditional deletion of *Trf2* (Fig. 1e). Telomeres that lack TRF2 undergo fusion mediated by classical non-homologous end joining[24–26]. In cells deficient in DNA ligase IV (LIG4), in which such telomere fusions are prevented[26], telomeres that lack TRF2 undergo 5′-end resection that is exacerbated by loss of 53BP1 or RIF1[8,19] (Fig. 1e). Similarly, the 5′-end resection was increased by REV7 or SHLD2 deficiency (Fig. 1f–h, Extended Data Fig. 4). When STN1 was depleted from cells that lack TRF2, resection at telomeres was significantly increased (Fig. 1f–h) and this effect was epistatic with REV7 (Fig. 1f–h). Thus, CST counteracts resection in a shieldin-dependent manner in the context of ATM signalling.

We next determined whether CST localized to damaged telomeres in a 53BP1- and shieldin-dependent manner. Myc-tagged CTC1 was detectable at telomeres with functional shelterin, whereas in cells that are deficient in POT1B—which show extended telomeric 3′ overhangs but no DNA damage signalling[27]—the localization of CTC1 at telomeres was minimal (Fig. 2a, b). When ATR was activated by deletion of *Tpp1* (Fig. 2a, right), CTC1 was again detectable at telomeres (Fig. 2a, b) despite the absence of POT1B. The recruitment of CTC1 to dysfunctional telomeres depended on ATR signalling, 53BP1 and shieldin (Fig. 2b, c). Similarly, Cre-mediated deletion of the single human POT1 protein from conditional *POT1*-knockout HT1080 cells[28] led to telomeric accumulation of STN1 that required ATR kinase (Fig. 2d–f). Thus, CST localizes to damaged telomeres in a shieldin-dependent manner.

Co-immunoprecipitation experiments showed that shieldin components could associate with CST (Fig. 2g, Extended Data Fig. 5a). In a yeast two-hybrid assay, CTC1 robustly interacted with SHLD1 and STN1 interacted with SHLD3 (Fig. 2h, Extended Data Fig. 5b). Weaker interactions were detectable between TEN1 and SHLD3; STN1 and SHLD1, SHLD2 or REV7; and CTC1 and REV7. Thus, shieldin binds CST through multiple direct interactions.

In human cells, STN1 co-localized with 53BP1 at DSBs induced by ionizing radiation (Fig. 3a, b) in a manner dependent on shieldin (Fig. 3b). Furthermore, STN1 was detectable at DSBs induced by FOKI in U2OS cells. This localization was diminished upon inhibition of ATM and ATR signalling, and required 53BP1 and shieldin (Fig. 3c–e, Extended Data Fig. 6a), which indicates that CST is recruited to sites of DNA damage by shieldin.

Because CST is associated with Polα–primase, we examined the localization of Polα at DSBs. Because Polα forms numerous S-phase foci (Extended Data Fig. 6b), we examined cells arrested in G2 (Fig. 3f,

[1]Laboratory for Cell Biology and Genetics, Rockefeller University, New York, NY, USA. [2]Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Brighton, UK. [3]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada. [4]Present address: Department of Medical and Health Sciences, Linköping University, Linköping, Sweden. [5]These authors contributed equally: Zachary Mirman, Francisca Lottersberger. *e-mail: delange@rockefeller.edu

© 2018 Springer Nature Limited. All rights reserved.

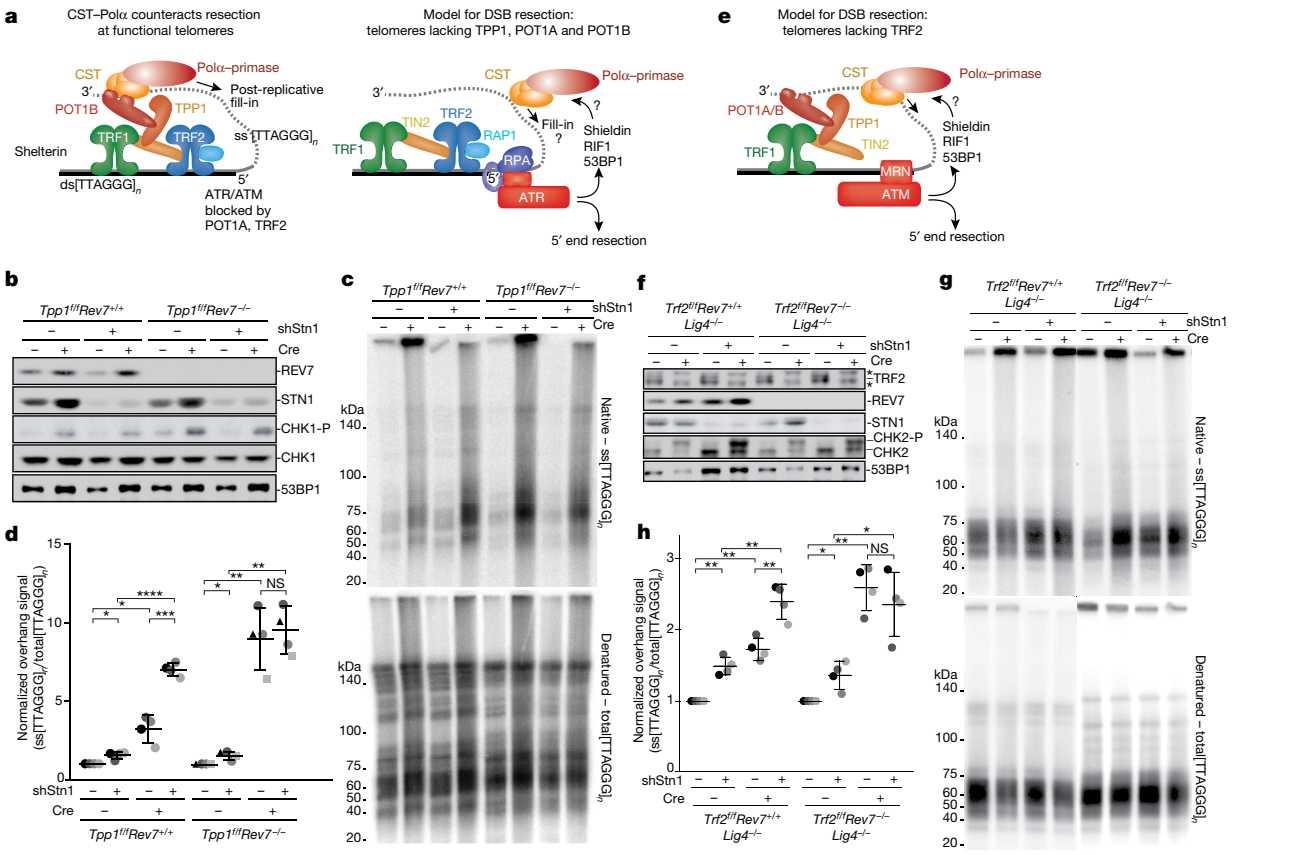


Fig. 1 | Shieldin and CST counteract resection at dysfunctional telomeres. a, Left, schematic showing POT1B-bound CST counteracting the resection of telomere ends. Right, depiction of telomeres that lack TPP1, POT1A and POT1B as a proxy for DSB resection. Telomeres that lack TPP1 undergo ATR-dependent hyper-resection that is repressed by 53BP1. **b**, Immunoblots showing loss of REV7 and STN1 in the indicated $Tpp1^{f/f}Rev7^{+/+}$ (*Rev7* is also known as *Mad2l2*) MEFs and $Tpp1^{f/f}Rev7^{-/-}$ (CRISPR) clones treated with Cre (96 h) and/or *Stn1* shRNA (shStn1) as indicated. CHK1-P serves as a proxy for TPP1 deletion. **c**, Quantitative analysis of telomere end resection in the cells shown in **b** using in-gel hybridization to detect the 3′ overhang (top), followed by rehybridization to the denatured DNA in the same gel (bottom) to determine the ratio of single-stranded (ss) TTAGGG to total TTAGGG signal. Representative of four experiments. **d**, Quantification of resection detected in **c**, determined from four independent experiments (different shades of grey) and showing mean and s.d. Three independent *Rev7* knockout clones were used (distinct symbols). **e**, Telomeres that lack TRF2 as a model for resection upon ATM activation. **f**, Immunoblots showing Cre-mediated deletion of TRF2 from $Trf2^{f/f}Lig4^{-/-}$ (*Trf2* is also known as *Terf2*) cells, CRISPR deletion of *Rev7*, shRNA-mediated reduction of STN1 and CHK2 phosphorylation. Asterisk, non-specific band. **g**, **h**, Telomere end resection analysis on the cells in **f**, as in **c**, **d**. Mean and s.d. from four independent experiments using two clones of each genotype. Note that the order of the samples is different in **h** when compared to **f**, **g**. All data panels in the figure are representative of four experiments. Mean indicated with centre bars and s.d. with error bars. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$, NS, not significant, two-tailed Welch's *t*-test.

Extended Data Fig. 6c). In cells that expressed HA-tagged STN1, Polα co-localized with STN1 at DSBs induced by FOKI (Fig. 3f, Extended Data Fig. 6c). The localization of Polα to DSBs was diminished upon inhibition of ATM and ATR signalling, and required 53BP1 and shieldin (Fig. 3f, Extended Data Fig. 6d), which demonstrates that Polα and CST require the same factors for their localization to DSBs.

Depletion of STN1 increased the percentage of cells that contained replication protein A (RPA) foci after exposure to ionizing radiation (Fig. 3g, i), increased the signal intensity of the RPA foci (Fig. 3h) and increased the overall RPA signal intensity per nucleus (Extended Data Fig. 7). Furthermore, deletion of CTC1 from a human HCT116 cell line[21] led to an increase in the phosphorylation of RPA upon irradiation (Fig. 3j) and CST depletion increased phosphorylation of RPA in irradiated MEFs (Fig. 3k). In cells that lack BRCA1, depletion of CST also increased the percentage of cells with RAD51 foci induced by ionizing radiation (Fig. 3l, m), which suggests that this depletion restores homology-directed repair. Conversely, on the basis of an assay for the fusion of telomeres that lack TRF2[26], it appears that the depletion of CST diminishes classical non-homologous end joining (Fig. 3n, o).

BRCA1-deficient cells become resistant to treatment with PARP1 inhibitors (PARPi) when 53BP1, RIF1 or shieldin are absent[3–9]. Similarly, STN1 or CTC1 depletion from $Brca1^{f/f}$ MEFs reduced the lethality of PARPi in BRCA1-deficient cells (Fig. 4a, b, Extended Data Fig. 8a–f). By contrast, in $Brca1^{f/f}$ subclones that lack 53BP1 or REV7, depletion of CTC1 or STN1 did not affect resistance to PARPi (Fig. 4c, Extended Data Fig. 8c–f). Furthermore, in BRCA1-deficient cells, CST depletion reduced the radial chromosomes induced by PARPi (Fig. 4d, e) and this effect was epistatic with 53BP1 and REV7 (Fig. 4e). These data are consistent with CST acting together with 53BP1 and shieldin to minimize formation of single-stranded DNA at DSBs.

To examine the consequences of Polα inhibition in PARPi-treated BRCA1-deficient cells without confounding S-phase effects, cells were arrested in G2 before addition of Polα inhibitors (Fig. 4f). Cells that experienced Polα inhibition in G2 showed reduced formation of radial chromosomes (Fig. 4f, Extended Data Fig. 8g). BrdU incorporation experiments confirmed that the mitotic cells we collected had passed through S phase during treatment with PARPi (Extended Data Fig. 8h–j). The effect of Polα inhibition with 10 μm CD437 was not exacerbated by depletion of CST (Fig. 4f). Collectively, these data are consistent with CST and Polα acting to limit the formation of recombinogenic 3′ overhangs at DSBs in BRCA1-deficient cells (Fig. 4g).

Our data suggest a sophisticated mechanism by which 53BP1 and shieldin act together with CST and Polα to fill in resected DSBs. At

© 2018 Springer Nature Limited. All rights reserved.

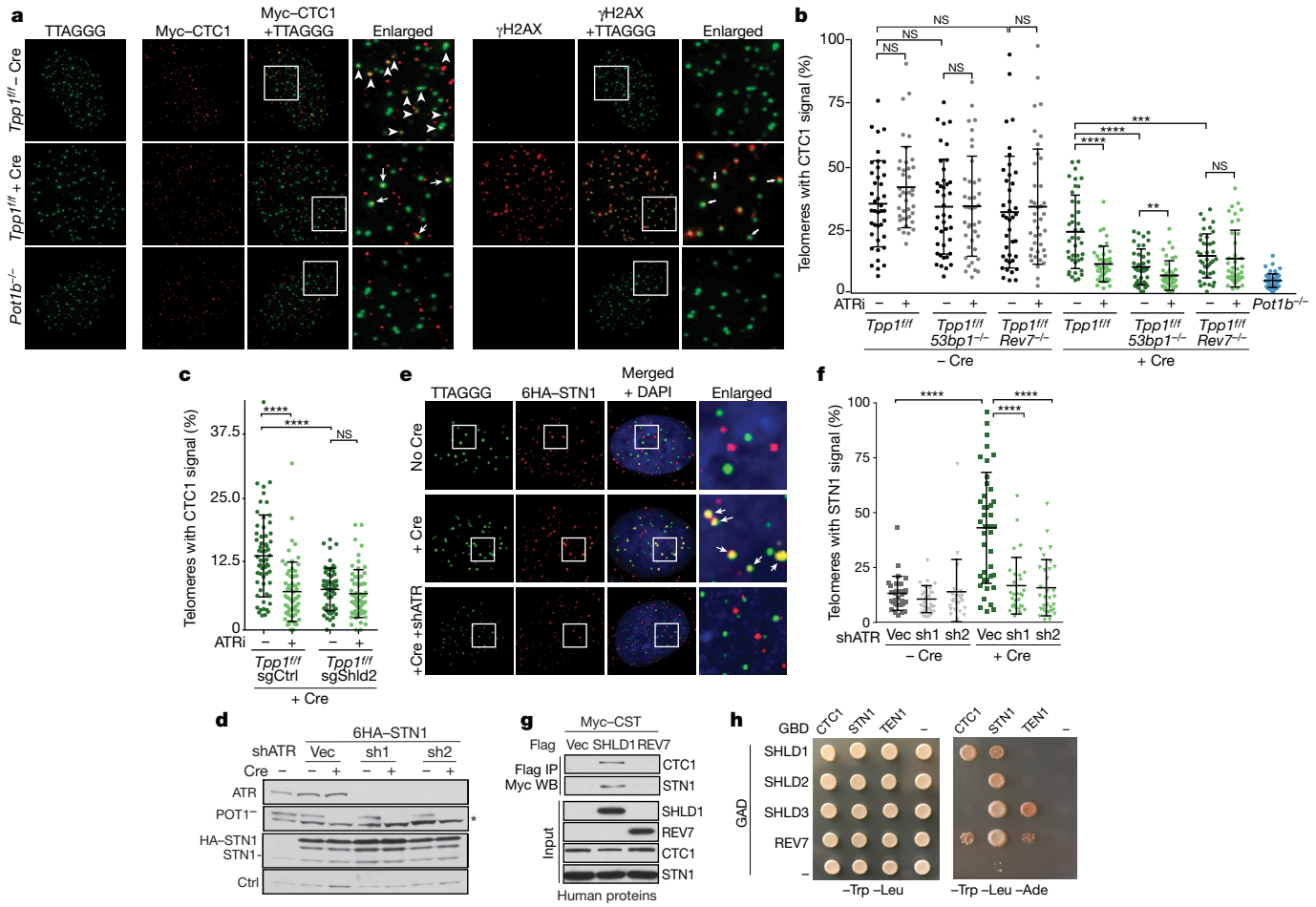


Fig. 2 | 53BP1- and shieldin-dependent localization of CST to dysfunctional telomeres. a, Left, representative immunofluorescence and fluorescence in situ hybridization (IF–FISH) for 6×-Myc-tagged CTC1 (red) at telomeres (false-coloured in green) in *Tpp1*$^{f/f}$ MEFs before and after Cre (96 h). Arrowheads, CTC1 at telomeres. *Pot1b*$^{-/-}$ cells control for spurious telomere–CTC1 co-localization. Right, the same nuclei showing γH2AX (red) at telomeres that lack TPP1. The γH2AX and CTC1 signals are both false-coloured in red. Arrows, telomeres with CTC1 and γH2AX. **b**, Quantification of the percentage of telomeres co-localizing with CTC1, detected as in **a**. Each dot represents one nucleus from the indicated *Tpp1*$^{f/f}$ cell lines with and without Cre and/or ATR inhibitor (ATRi). *53bp1* is also known as *Trp53bp1*. Mean and s.d. from three independent experiments. **c**, As in **b**, but using *Tpp1*$^{f/f}$ cells treated with a *Shld2* single-guide RNA (sgRNA) (sgShld2) or a control sgRNA (sgCtrl). Mean and s.d. as in **b**. **d**, Immunoblots for POT1 deletion, ATR knockdown and HA-tagged STN1 (6HA–STN1) in conditional *POT1*-knockout human HT1080 cells. Asterisk, non-specific band. sh1 and sh2 denote two different *ATR* shRNAs (see Methods for details). **e**, IF–FISH showing telomeric DNA co-localizing with STN1 in cells as in **d**, treated with Cre (96 h) and *ATR* shRNAs. **f**, Quantification of STN1 localization at telomeres before and after *POT1* deletion, with or without *ATR* shRNAs as in **e**. Mean and s.d. from three independent experiments. Each symbol represents one nucleus. **g**, Immunoprecipitation of human CST (each subunit Myc-tagged) with Flag-tagged human SHLD1 or REV7 co-expressed in 293T cells. **h**, Yeast two-hybrid assay for interaction between CST and shieldin subunits. All data panels in the figure are representative of three experiments. Mean indicated with centre bars and s.d. with error bars. GAD, Gal4 activation domain; GBD, Gal4 DNA-binding domain; Vec, empty vector. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$, two-tailed Welch's *t*-test.

telomeres, the POT1–TPP1 heterodimer recruits CST–Polα–primase to fill in part of the 3′ overhang formed after telomere end resection (Fig. 4g). We propose that at sites of DNA damage, shieldin recruits CST–Polα–primase for the purpose of filling in resected DSBs. In both settings, CST is tethered, allowing CST to engage single-stranded DNA despite its modest affinity for this substrate[29] and enabling regulation of the fill-in reaction through recruitment. Recent data have shown that 53BP1 represses mutagenic single-strand annealing, possibly by preventing excessive resection[30]. Our findings regarding CST–Polα could explain this observation. At telomeres, partial fill-in by CST–Polα counteracts hyper-resection but leaves a 3′ overhang that can form a t-loop, a process similar to the initiation of homology-directed repair[12,13]. At DSBs, CST–Polα could similarly counteract hyper-resection—and thus single-strand annealing—and generate a 3′ overhang that is sufficient for homology-directed repair. In BRCA1-deficient cells, this fill-in reaction, together with the persistence of CST–shieldin at the DSBs, could block homology-directed repair and result in lethal mis-repair.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0324-7.

Received: 2 January 2018; Accepted: 10 May 2018;
Published online 18 July 2018.

1. Zimmermann, M. & de Lange, T. 53BP1: pro choice in DNA repair. *Trends Cell Biol.* **24**, 108–117 (2014).
2. Panier, S. & Boulton, S. J. Double-strand break repair: 53BP1 comes into focus. *Nat. Rev. Mol. Cell Biol.* **15**, 7–18 (2014).
3. Bouwman, P. et al. 53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers. *Nat. Struct. Mol. Biol.* **17**, 688–695 (2010).
4. Xu, G. et al. REV7 counteracts DNA double-strand break resection and affects PARP inhibition. *Nature* **521**, 541–544 (2015).
5. Bunting, S. F. et al. 53BP1 inhibits homologous recombination in *Brca1*-deficient cells by blocking resection of DNA breaks. *Cell* **141**, 243–254 (2010).

© 2018 Springer Nature Limited. All rights reserved.

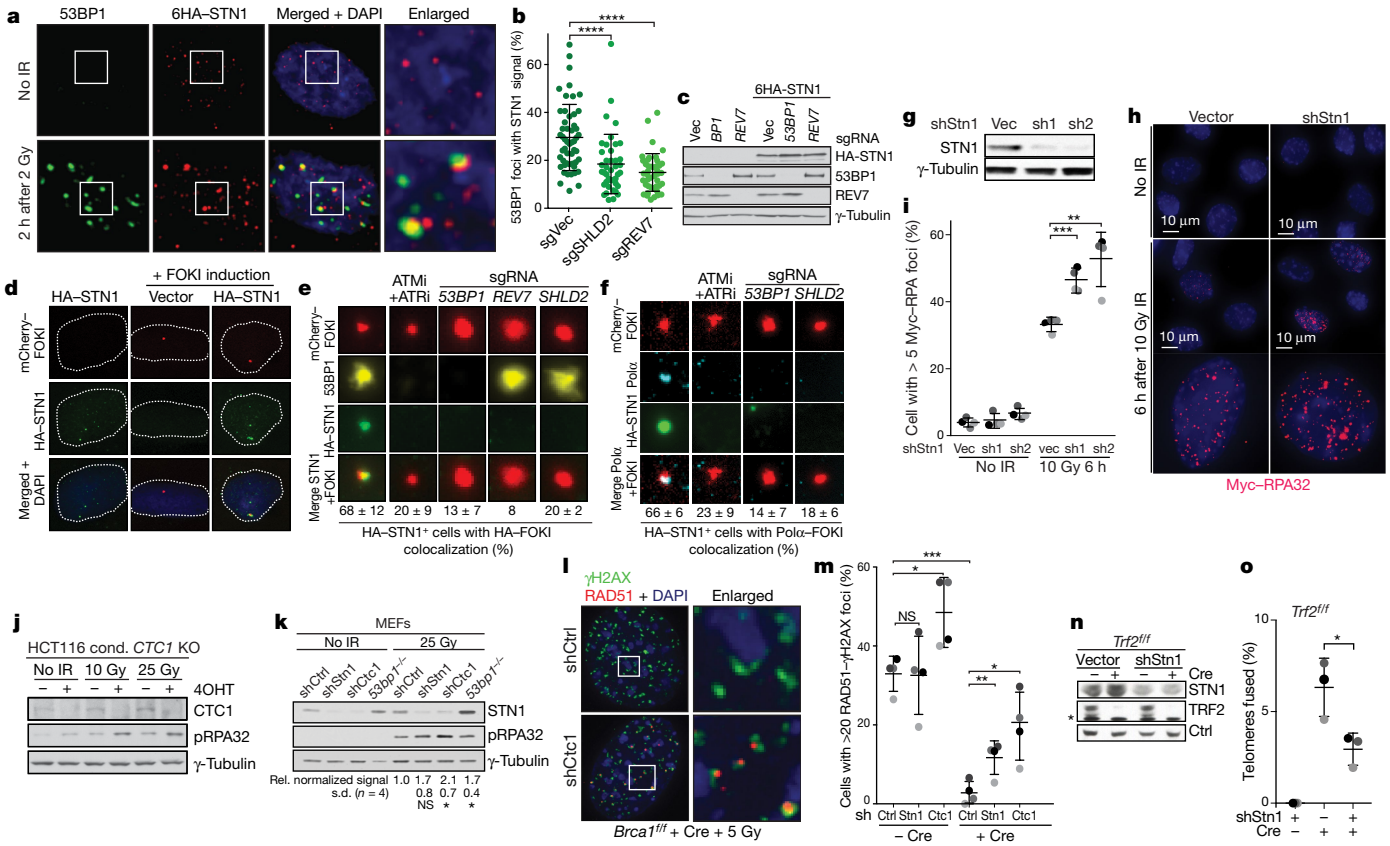


Fig. 3 | CST localizes to DSBs and represses formation of single-stranded DNA. a, Immunofluorescence for 53BP1 and HA–STN1 in HT1080 cells treated with ionizing radiation. **b**, Quantification of 53BP1–STN1 co-localization as in **a**, in cells with the indicated sgRNAs. Mean and s.d. from three independent experiments, >15 nuclei per experiment for each experimental setting. **c**, Immunoblots for the indicated proteins in FOKI–LacI U2OS cells treated with the indicated sgRNAs. *53BP1* is also known as *TRP53BP1*. **d**, Immunofluorescence for mCherry–FOKI (red) and HA–STN1 (green) in FOKI–LacI U2OS cells as in **c**. **e**, Examples of HA–STN1 co-localizing with FOKI foci in cells as in **d**, treated with ATM inhibitor (ATMi) and ATRi, or the indicated sgRNAs and quantification of STN1–FOKI co-localization. Mean and s.d. from three independent experiments with >80 induced nuclei analysed for each condition. **f**, As in **e**, but monitoring Polα at DSBs in G2-arrested cells that express HA–STN1. **g**, Immunoblot for STN1 knockdown in MEFs that express Myc–RPA32. sh1 and sh2 denote two different *Stn1* shRNAs (see Methods for details). **h**, Immunofluorescence for Myc–RPA32 after 10 Gy ionizing radiation (6 h) in MEFs. **i**, Quantification of cells with RPA foci as in **h**, in >30 nuclei for each condition in four independent experiments (grey shading) with mean and s.d. **j**, **k**, Immunoblots for ionizing radiation-induced RPA phosphorylation (pRPA32, phosphorylated at Ser4 and Ser8) after deletion of *CTC1* from human cells (**j**) or after depletion of STN1, CTC1 or 53BP1 from MEFs (**k**). **l**, Immunofluorescence for RAD51–γH2AX co-localization at DSBs induced by ionizing radiation, in BRCA1-deficient cells treated with *Ctc1* shRNA. **m**, Quantification of data in **l**. Mean and s.d. from four independent experiments (grey shading), >60 nuclei per experiment. **n**, Immunoblot for STN1 knockdown and TRF2 deletion from $Trf2^{f/f}Rosa^{creER}$ MEFs. Asterisk denotes non-specific band. **o**, Effect of STN1 knockdown on telomere–telomere fusions. Mean and s.d. from three independent experiments with >6,000 telomeres each. All immunofluorescence and immunoblots shown are representative of three experiments. Mean indicated with centre bars and s.d. with error bars. IR, ionizing radiation; KO, knockout. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$, two-tailed Welch's *t*-test.

6. Chapman, J. R. et al. RIF1 is essential for 53BP1-dependent nonhomologous end joining and suppression of DNA double-strand break resection. *Mol. Cell* **49**, 858–871 (2013).
7. Boersma, V. et al. MAD2L2 controls DNA repair at telomeres and DNA breaks by inhibiting 5′ end resection. *Nature* **521**, 537–540 (2015).
8. Zimmermann, M., Lottersberger, F., Buonomo, S. B., Sfeir, A. & de Lange, T. 53BP1 regulates DSB repair using Rif1 to control 5′ end resection. *Science* **339**, 700–704 (2013).
9. Noordermeer, S. M. et al. The shieldin complex mediates 53BP1-dependent DNA repair. *Nature* https://doi.org/10.1038/s41586-018-0340-7 (2018).
10. Price, C. M. et al. Evolution of CST function in telomere maintenance. *Cell Cycle* **9**, 3177–3185 (2010).
11. Casteel, D. E. et al. A DNA polymerase-α·primase cofactor with homology to replication protein A-32 regulates DNA replication in mammalian cells. *J. Biol. Chem.* **284**, 5807–5818 (2009).
12. Griffith, J. D. et al. Mammalian telomeres end in a large duplex loop. *Cell* **97**, 503–514 (1999).
13. Doksani, Y., Wu, J. Y., de Lange, T. & Zhuang, X. Super-resolution fluorescence imaging of telomeres reveals TRF2-dependent T-loop formation. *Cell* **155**, 345–356 (2013).
14. Palm, W. & de Lange, T. How shelterin protects mammalian telomeres. *Annu. Rev. Genet.* **42**, 301–334 (2008).
15. Lazzerini-Denchi, E. & Sfeir, A. Stop pulling my strings–what telomeres taught us about the DNA damage response. *Nat. Rev. Mol. Cell Biol.* **17**, 364–378 (2016).
16. Wu, P., Takai, H. & de Lange, T. Telomeric 3′ overhangs derive from resection by Exo1 and Apollo and fill-in by POT1b-associated CST. *Cell* **150**, 39–52 (2012).
17. Goulian, M., Heard, C. J. & Grimm, S. L. Purification and properties of an accessory protein for DNA polymerase α/primase. *J. Biol. Chem.* **265**, 13221–13230 (1990).
18. Sfeir, A. & de Lange, T. Removal of shelterin reveals the telomere end-protection problem. *Science* **336**, 593–597 (2012).
19. Lottersberger, F., Bothmer, A., Robbiani, D. F., Nussenzweig, M. C. & de Lange, T. Role of 53BP1 oligomerization in regulating double-strand break repair. *Proc. Natl Acad. Sci. USA* **110**, 2146–2151 (2013).
20. Kibe, T., Zimmermann, M. & de Lange, T. TPP1 blocks an ATR-mediated resection mechanism at telomeres. *Mol. Cell* **61**, 236–246 (2016).
21. Feng, X., Hsu, S. J., Kasbek, C., Chaiken, M. & Price, C. M. CTC1-mediated C-strand fill-in is an essential step in telomere length maintenance. *Nucleic Acids Res.* **45**, 4281–4293 (2017).
22. Gu, P. et al. CTC1 deletion results in defective telomere replication, leading to catastrophic telomere loss and stem cell exhaustion. *EMBO J.* **31**, 2309–2321 (2012).
23. Lottersberger, F., Karssemeijer, R. A., Dimitrova, N. & de Lange, T. 53BP1 and the LINC complex promote microtubule-dependent DSB mobility and DNA repair. *Cell* **163**, 880–893 (2015).
24. van Steensel, B., Smogorzewska, A. & de Lange, T. TRF2 protects human telomeres from end-to-end fusions. *Cell* **92**, 401–413 (1998).
25. Karlseder, J., Broccoli, D., Dai, Y., Hardy, S. & de Lange, T. p53- and ATM-dependent apoptosis induced by telomeres lacking TRF2. *Science* **283**, 1321–1325 (1999).

© 2018 Springer Nature Limited. All rights reserved.

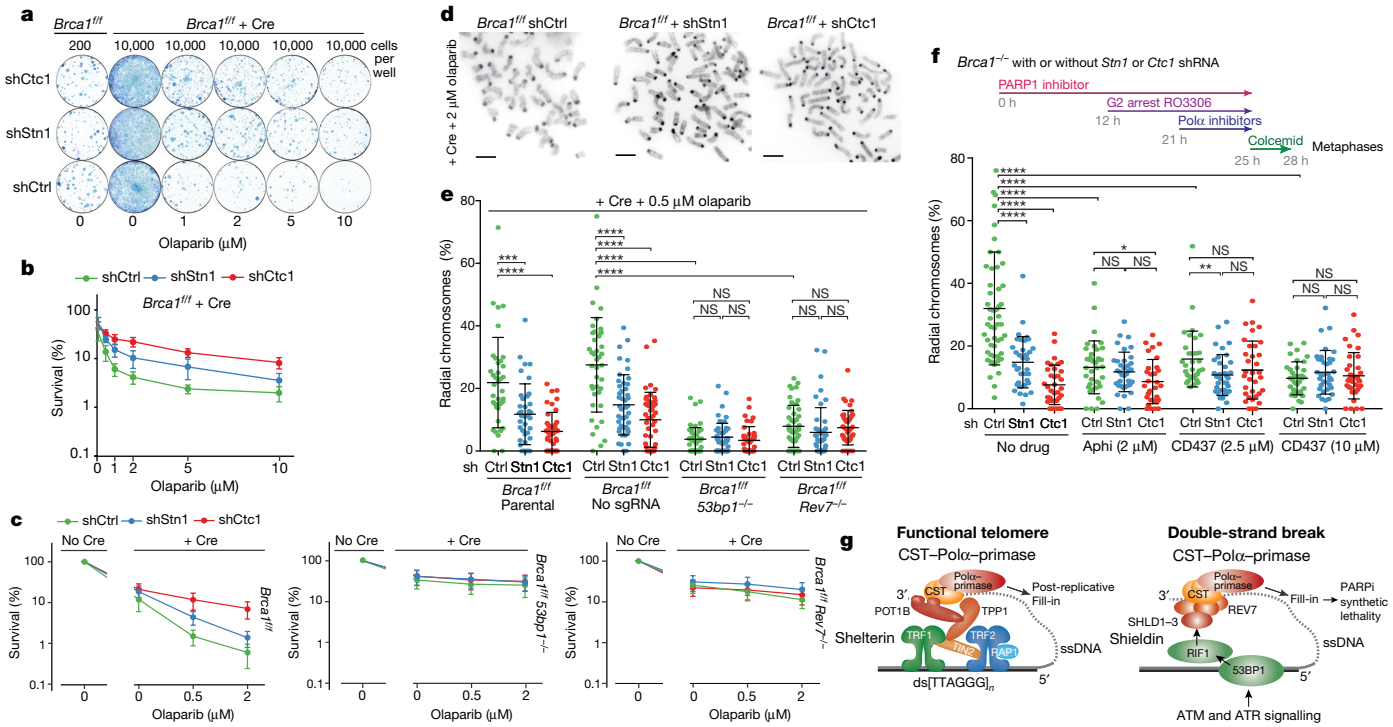


Fig. 4 | CST and Polα affect the outcome of PARPi in BRCA1-deficient cells. a, Colonies detected in a PARPi survival assay using *Brca1*$^{f/f}$ MEFs with or without Cre and *Ctc1* or *Stn1* shRNAs. **b**, Graphical representation of data in **a** from three independent experiments. **c**, Epistasis analysis of PARPi resistance induced by the absence of 53BP1 or REV7, and depletion of CST subunits. Mean (symbol) and s.e.m. (error bars) from three independent experiments. **d**, PARPi-induced radial chromosomes in BRCA1-deficient cells. Scale bars, 1 μm. **e**, Mean (centre bar) and s.d. (error bars) of percentage of misrejoined (radial) chromosomes in >10 metaphases per experimental setting for each of three independent experiments. Each dot represents one metaphase. **f**, Effect of Polα inhibition on radial formation in PARPi-treated *Brca1*$^{-/-}$ cells using the experimental timeline shown. Mean (centre bar) and s.d. (error bars) of the percentage of radial chromosomes in >10 metaphases per experimental setting for each of three independent experiments. Each dot represents one metaphase. Aphi, aphidicolin. **g**, Graphical representation of the similar mechanisms by which resection is counteracted at functional telomeres and at DSBs. **a** and **d** are representative of three experiments. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$, two-tailed Welch's *t*-test.

26. Celli, G. B. & de Lange, T. DNA processing is not required for ATM-mediated telomere damage response after TRF2 deletion. *Nat. Cell Biol.* **7**, 712–718 (2005).
27. Hockemeyer, D., Daniels, J. P., Takai, H. & de Lange, T. Recent expansion of the telomeric complex in rodents: two distinct POT1 proteins protect mouse telomeres. *Cell* **126**, 63–77 (2006).
28. Takai, H. et al. A POT1 mutation implicates defective telomere end fill-in and telomere truncations in Coats plus. *Genes Dev.* **30**, 812–826 (2016).
29. Hom, R. A. & Wuttke, D. S. Human CST prefers G-rich but not necessarily telomeric sequences. *Biochemistry* **56**, 4210–4218 (2017).
30. Ochs, F. et al. 53BP1 fosters fidelity of homology-directed DNA repair. *Nat. Struct. Mol. Biol.* **23**, 714–721 (2016).

Acknowledgements We thank D. White for mouse husbandry; N. Bosco, R. Karssemeijer, L. Timashev and Y. Doksani for help with CRISPR gene knockouts, image analysis and generating MEFs; and R. Greenberg and C. Price for providing cell lines. The Rockefeller University BioImaging Center provided assistance. This work was supported by grants from the NCI (R35CA210036), ACS and BCRF to T.d.L., a grant from the CIHR (FDN143343) to D.D. and the Banting Postdoctoral fellowship to M.Z.

Author contributions M.Z. initiated this work in the de Lange laboratory. T.K. and F.L. performed resection assays. Z.M. and H.T. performed CST and Polα localization assays and RPA phosphorylation assays. Z.M. performed PARPi and RAD51 assays. Y.G. and T.K. analysed RPA foci. A.B. performed yeast two-hybrid assays. K.T. and H.T. performed co-immunoprecipitation analysis. K.T. and T.d.L. performed telomere fusion assays. D.D. provided information, reagents and advice. T.d.L. conceived the study and wrote the paper with input from all co-authors.

Competing interests D.D. is a founder of, owns equity in and receives funding from Repare Therapeutics.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0324-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0324-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture and expression constructs. *Brca1*$^{f/f}$ and *Trf2*$^{f/f}$*Lig4*$^{-/-}$ MEFs were derived from *Brca1*$^{f/f}$ mice[31], *Trf2*$^{f/f}$ mice[26] and *Lig4*$^{+/-}$ mice[32] by standard crosses. Mice were housed and cared for under Rockefeller University IACUC protocol 16865-H at the Rockefeller University's Comparative Bioscience Center, which provides animal care according to NIH guidelines. MEFs were isolated from embryonic-day 12.5 embryos and immortalized with pBabeSV40 large T antigen (a gift from G. Hannon) at early passage (P2/3), as previously described[26]. Genotypes were determined by Transnetyx, using real time PCR with allele-specific probes. *Tpp1*$^{f/f}$, *Tpp1*$^{f/f}$*53bp1*$^{-/-}$ (ref. [33]), *Tpp1*$^{f/f}$*Rif1*$^{f/f}$ or *Tpp1*$^{f/f}$*Rif1*$^{f/+}$ (ref. [20]) and *Pot1b*$^{Stop/Stop}$ (ref. [27]) MEFs have previously been described. MEFs and U2OS cells were cultured in Dulbecco's modified Eagle medium (DMEM, Corning) supplemented with 15% fetal bovine serum (FBS) (Gibco), non-essential amino acids (Gibco), 2 mM L-glutamine (Gibco), 100 U/ml penicillin, 100 μg/ml streptomycin (Gibco), 50 μM β-mercaptoethanol (Sigma). 293T, Phoenix and conditional *POT1*-knockout HT1080 clone c5[28] cells were cultured in DMEM supplemented with 10% bovine calf serum (BCS), non-essential amino acids, L-glutamine and penicillin–streptomycin as above. For most Cre-mediated gene deletion experiments (see exceptions below), retroviral infections with pMMP Hit & Run Cre were repeated three times[26]. Time points of cell collection indicate hours after the second Cre infection.

U2OS cells containing a LacO array and a tamoxifen- and Shield1-regulated mCherry–FOKI–LacI fusion were used as described[34]. Cells were collected 4 h after induction of FOKI by addition of 0.1 μM Shield1 and 10 μg/ml 4-OHT. Human *CTC1*$^{f/f}$ HCT116 cells[21] were cultured in McCoy's 5A supplemented with 10% FCS, non-essential amino acids, L-glutamine and penicillin–streptomycin as above. *CTC1* gene deletion was induced with 0.5 μM 4-OHT for 5 h. Gene deletion was confirmed by western blot using anti-CTC1 antibody (MABE1103, Millipore).

Mouse CTC1 tagged at the N terminus with a 6×Myc tag was delivered by retroviral transduction using pLPC or pWZL retroviral vectors. Human STN1 tagged at the N terminus with a 6×HA tag was delivered using the pLPC vector. Myc-tagged RPA32[35], and 53BP1 wild type and 53BP1ΔRif1[23] constructs were as described. Retroviral gene delivery was performed as described[16].

RNA depletion with shRNAs in pLKO.1 (Open Biosystems) was performed using the following shRNA target sites: shStn1 1: 5′-GATCCTGTGTTTCTAGCCTTT-3′ (TRCN0000180836, Sigma); shStn1 2: 5′-GCTGTCATCAGCGTGAAAGAA-3′ (TRCN0000184261, Sigma); shCtc1: 5′-CGGCAGATCACAGCATGATAA-3′; shAtr 1: 5′-CTGTGGTTGTATCTGTTCAAT-3′ (TRCN0000039613, Sigma); shAtr 2: 5′-GATGAACACATGGGATATTTA-3′ (TRCN0000196538, Sigma). Lentiviral constructs were co-transfected with packaging vectors into 293T cells and cells infected with the viral supernatant were selected in puromycin as described[16].

Drug treatments were as follows. ATR inhibition: 2.5 μM ETP-46464 (Sigma), 24 h; PARP1 inhibition: 0.1–10 μM Olaparib (Selleck Chemicals), 24 h; G2 arrest: 9 μM RO-3306 (Sigma), 12 h; polymerase α inhibition: 2.5 or 10 μM CD437 (Sigma) or 2 μM Aphidicolin, 4 h. ATM and ATR inhibition: 10 μM KU55933 (Selleck Chemical) with ATR inhibition as above for 4 h during induction of FOKI nuclease.

CRISPR–Cas9 gene disruption. Clonal cell lines with disruption of mouse 53BP1 were generated using Cas9 vector (Addgene) and sgRNA (sg53bp1 (2), 5′-GAGAATCTTCTATTATC-(PAM)-3′; sg53bp1 (3), 5′- GCATCTGCA GATTAGGA-(PAM)-3′[20]) delivered by nucleofection (Amaxa Kit R, Lonza). Clones were screened by immunoblotting and bi-allelic gene disruption was verified by Sanger sequencing of Topo-cloned PCR products of the relevant locus (sequences available on request). Clonal cell lines with mouse *Rev7* gene disruption were isolated similarly using the following sgRNAs: sgRev7(2), 5′-GTGTCCCCACCACAGTGG-(PAM)-3′; and sgRev7(3), 5′-GCCGGTTC AGGTGAGCCC-(PAM)-3′) (disrupted gene sequences available on request). Oligonucleotides were purchased from Sigma-Aldrich and cloned into the AflII-digested gRNA expression vector (Addgene) by Gibson Assembly (New England Biolabs). For isolation of populations with CRISPR–Cas9 disruption of mouse *Shld2* (FAM35A), 293T cells were transfected with lentiCrispr-v2-Shld2-sgRNA (5′- ATCAGTCAGATCCCTGCGTT-(PAM)-3′) or the vector control. The lentiviral supernatant was used for infection of *Tpp1*$^{f/f}$ or *Trf2*$^{f/f}$*Lig4*$^{-/-}$ MEFs; infections were done six times at 6–12 h intervals. Infected cells were then selected in puromycin for 3–5 days before Cre infection.

FOKI–LacI U2OS cells were infected with the 6×-HA-tagged human STN1 retrovirus and selected in puromycin. Subsequently, cells were subjected to lentiviral infection with lentiCrispr-v2 carrying sgRNA for human *53BP1*, *SHLD2* or *REV7* and selected in blasticidin for 3 days. Target sequences for gene disruption are as follows: human *53BP1* sgRNA1 (5′-CAGAATCATCCTCTAGAACC-(PAM)-3′), *53BP1* sgRNA2 (5′-TTGATCTCACTTGTGATTCG-(PAM)-3′), *SHLD2* sgRNA1 (5′-TCTGGAGAACCAATAGATTC-(PAM)-3′), *SHLD2* sgRNA2 (5′-TTTGAGCTAAAAAAGCAACC-(PAM)-3′), *REV7* sgRNA1 (5′-CCT CAACTTTGGCCAAGGTA-(PAM)-3′), *REV7* sgRNA2 (5′-TATACTGATTC AGCTCCGGG-(PAM)-3′). For each gene the two sgRNAs were either used individually or together.

In-gel analysis of single-stranded telomeric DNA. Mouse telomeric overhang and telomeric restriction fragment patterns were analysed 96–120 h after Cre treatment by in-gel hybridization with a γ-^{32}P-ATP end-labelled $(AACCCT)_4$ probe, as previously described[26]. Treatment with *E. coli* exonuclease I before MboI digestion was used to verify the 3′ terminal position of the single-stranded DNA as previously described[20]. ImageQuant software was used to quantify the single-stranded telomere overhang signals and the signal from total telomeric DNA in the same lane in the denatured gel. In each experiment, this ratio was set to 1 for lanes not treated with Cre or shRNA and the ratios for the treated samples are given relative to this control.

Flow cytometry. FACS was performed as previously described[36] with gating.

Immunoblotting. Immunoblotting was performed as described[16] with the following antibodies: 53BP1 (175933, Abcam; NB100-304, Novus Biological); ATR (sc-1887, Santa Cruz Biotechnology); BRCA1 (MAB22101, R+D systems); CHK1 (sc-8408, Santa Cruz Biotechnology); CHK1-S345-P (#2341S; Cell Signaling Technology); CHK2 (BD 611570, BD Biosciences); Flag-tag (M2, Sigma; F1804, Sigma); γ-tubulin (GTU488, Sigma); MAD2L2/REV7 (ab180579, Abcam); Myc-tag (9B11, Cell Signaling Technology); OBFC1/STN1 (E10-376450, Santa Cruz Biotechnology); Tagged TEN1 was not detectable by immunoblotting of transfected 293T cells.

For detection of RPA phosphorylation, conditional *CTC1* HCT116 cells or MEFs were irradiated and collected 3 h later. Cells were washed in PBS, and then collected by scraping in Laemmli sample buffer, boiling for 5 min and shearing through a syringe. Proteins were separated by SDS–PAGE on 8–16% Tris-glycine gradient gels (Invitrogen), and transferred to nitrocellulose overnight. Immunoblotting for pRPA followed standard protocols with blocking in 5% milk/TBST and the pRPA antibody (S4/S8; Bethyl) diluted 1:1,000 in 1% milk/TBST.

Immunoprecipitation. Immunoprecipitation was carried out as described[16]. The following plasmids were used: pLPC-flag-POT1a and pLPC-flag-POT1b[27]; pLPC-myc-mouse Ctc1, pLPC-myc-mouse Stn1 and pLPC-myc-mouse Ten1[16]; pLPC-myc-human Ctc1, pLPC-myc-human Stn1, pLPC-myc-human Ten1, pCD-NA5-flag-human Shld1 (C20orf196) and pLPC-myc-hRev7, pLPC-flag-mouse Shld1 (orthologue of C20orf196). Human *REV7*, *SHLD1*, *CTC1*, *STN1* and *TEN1* ORFs were generated by PCR and mouse *Shld1* was generated by RT–PCR. Co-transfection of *Shld1* and *Rev7* with CST in 293T cells was performed using calcium phosphate co-precipitation. Lysates were prepared in lysis buffer containing 50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 10% glycerol and 0.1% NP-40, Complete protease inhibitor mix (Roche), and PhosSTOP phosphatase inhibitor mix (Roche) and 50 U benzonase.

Yeast two-hybrid assays. For yeast two-hybrid analysis, full-length versions of human CST and shieldin components were cloned into the NdeI site of the pGBKT7 and pGADT7 vectors (Clontech). Plasmids in the indicated pairwise combinations were co-transformed into budding yeast strain PJ69-4A (*MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4 gal80 LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ*) and selected on synthetic complete drop-out medium lacking tryptophan and leucine. Protein interactions were tested by plating on the same medium but also lacking adenine.

Immunofluorescence and IF–FISH. Previously published procedures were followed for immunofluorescence and IF–FISH[16]. Immunofluorescence for Myc-tagged RPA32 or CTC1 (mouse monoclonal, 9B11 or rabbit monoclonal, 71D10, Cell Signaling Technology), HA-tagged STN1 (3724, Cell Signaling Technology), endogenous Polα (sc-137021, Santa Cruz), and 53BP1 (612522, BD Biosciences) was carried out using the cytoskeleton extraction protocol[37]. Intensity measurements of RPA32–Myc immunofluorescence were performed in FIJI as follows: nuclei were identified using thresholding, segmented and identified as regions of interest. The average image background was then subtracted from the image, and the total raw pixel intensity within each area of interest in the channel of interest was calculated. RAD51 (70-001, Bioacademia), and γH2AX (05636, Millipore) were detected in cells fixed in 3% PFA, and foci showing co-localization of RAD51 with γH2AX were quantified. Immunofluorescence imaging was performed on a Zeiss Axioplan II microscope equipped with a Hamamatsu C4742-95 camera using Volocity software or on a DeltaVision (Applied Precision) equipped with a cooled charge-coupled device camera (DV Elite CMOS Camera), a PlanApo 60× 1.42 NA objective or 100× 1.40 NA objective (Olympus America), and SoftWoRx software.

Telomere fusion assays. SV40LT-immortalized *Trf2*$^{f/f}$*Rosa*^{creER} cells were infected with *Stn1* shRNA (or the empty vector) and 24 h later Cre was induced for 24 h with 4-OHT. Cells were collected, counted (to rule out a proliferation defect) and processed for telomeric FISH on metaphases 72 h after Cre induction. This early time point was selected to avoid any effect of the *Stn1* shRNA on proliferation, as

© 2018 Springer Nature Limited. All rights reserved.

diminished proliferation reduces fusion frequencies. Telomere fusions were scored as previously described[26].

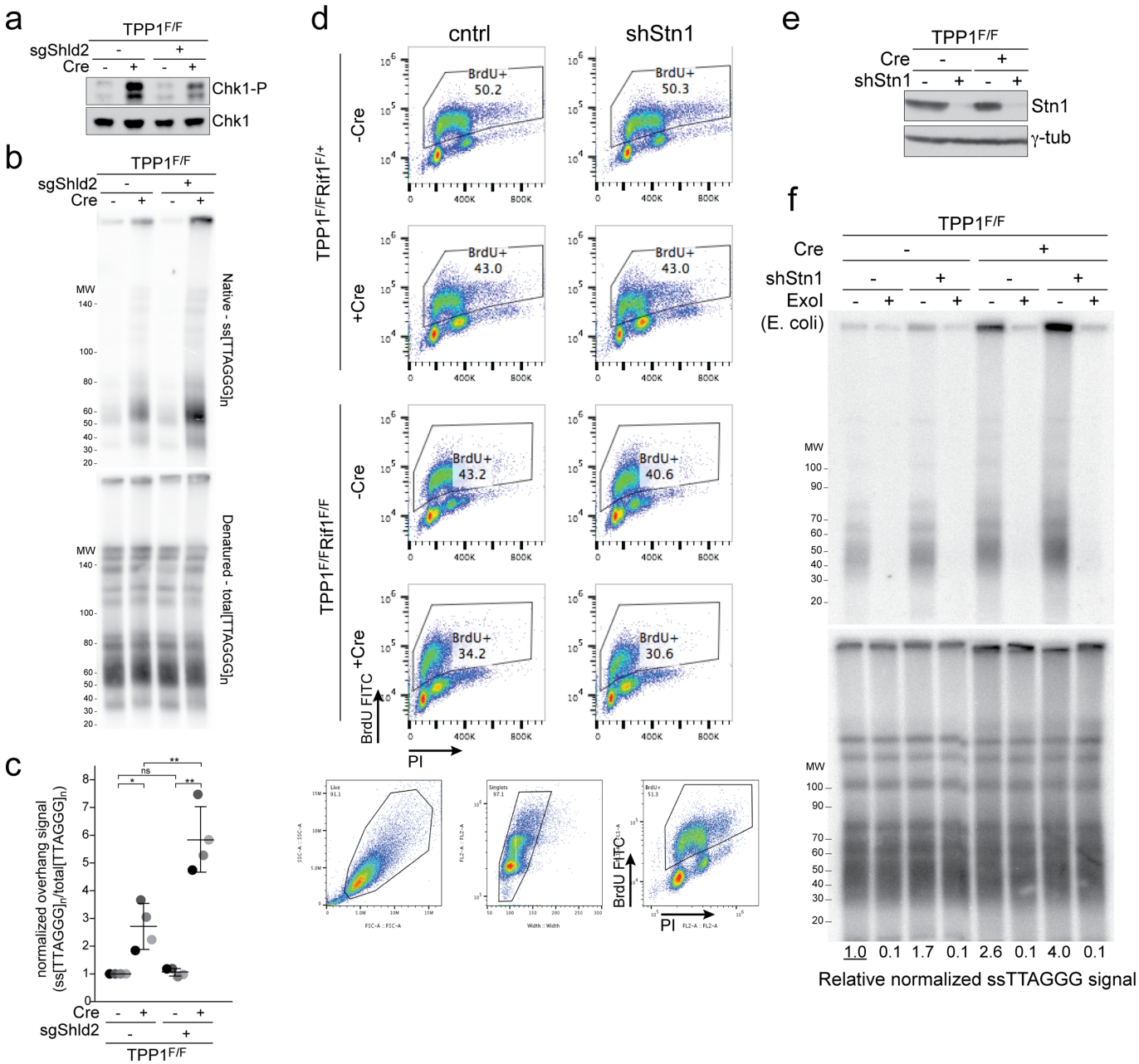
Survival assays and chromosome analysis. PARPi survival assays and analysis of misrejoined chromosomes were carried out as described[19], except that for analysis of radial chromosomes, MEFs were incubated with 0.5 μM Olaparib (AZD2281) for 24 h before collection. For the survival assays, MEFs were seeded in 6-well plates in duplicate at 10, 50, 100, 500, 1,000, 5,000 or 10,000 cells per well. After 24 h, cells were treated with Olaparib at the indicated concentrations for 24 h. Cells were then provided with medium without Olaparib and incubated for one week with a medium change at day 4. Colonies were fixed and stained with 50% methanol, 2% methylene blue, rinsed with water and dried before counting. The survival percentage at each PARPi concentration compared to untreated cells was calculated using wells with 10–100 colonies. Two technical replicates at two cell concentrations were scored for each condition in three independent experiments.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All data that were generated and/or analysed in this study are included in the published paper and its Supplementary Information.

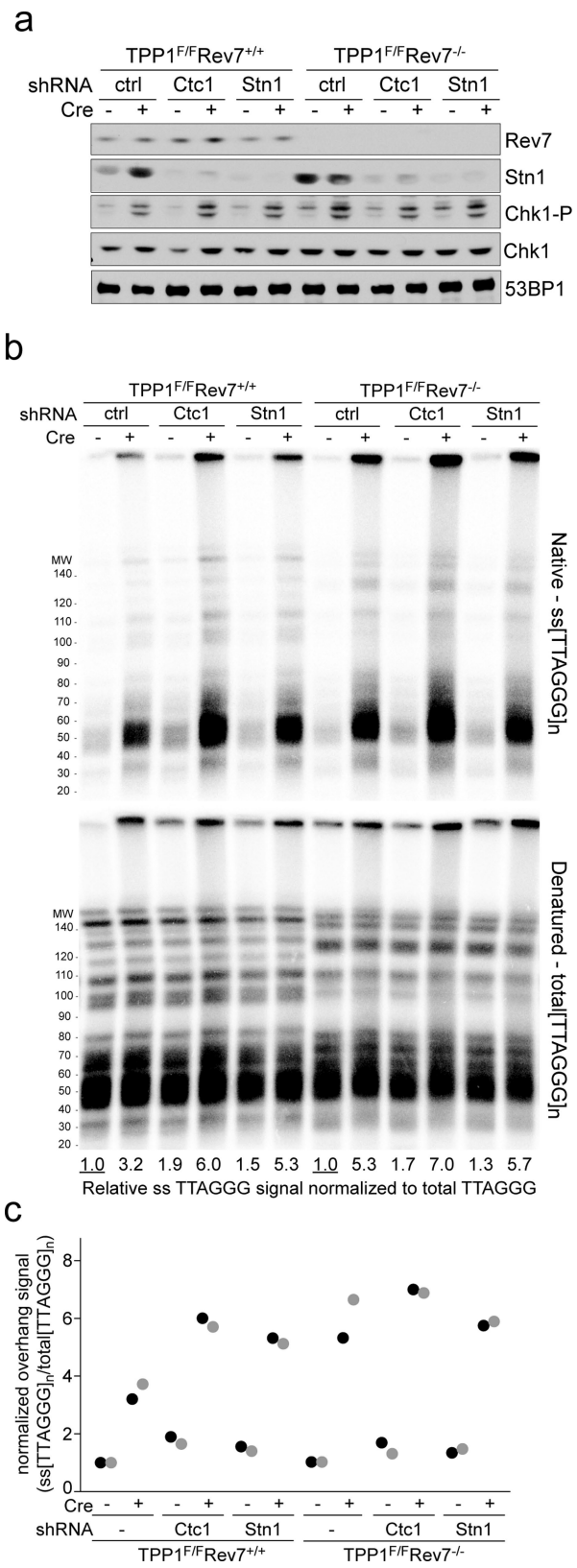
31. Xu, X. et al. Centrosome amplification and a defective G2-M cell cycle checkpoint induce genetic instability in BRCA1 exon 11 isoform-deficient cells. *Mol. Cell* **3**, 389–395 (1999).
32. Frank, K. M. et al. Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature* **396**, 173–177 (1998).
33. Kibe, T., Osawa, G. A., Keegan, C. E. & de Lange, T. Telomere protection by TPP1 is mediated by POT1a and POT1b. *Mol. Cell. Biol.* **30**, 1059–1066 (2010).
34. Tang, J. et al. Acetylation limits 53BP1 association with damaged chromatin to promote homologous recombination. *Nat. Struct. Mol. Biol.* **20**, 317–325 (2013).
35. Gong, Y., Handa, N., Kowalczykowski, S. C. & de Lange, T. PHF11 promotes DSB resection, ATR signaling, and HR. *Genes Dev.* **31**, 46–58 (2017).
36. Takai, H., Wang, R. C., Takai, K. K., Yang, H. & de Lange, T. Tel2 regulates the stability of PI3K-related protein kinases. *Cell* **131**, 1248–1259 (2007).
37. Mirzoeva, O. K. & Petrini, J. H. DNA damage-dependent nuclear dynamics of the Mre11 complex. *Mol. Cell. Biol.* **21**, 281–288 (2001).

© 2018 Springer Nature Limited. All rights reserved.



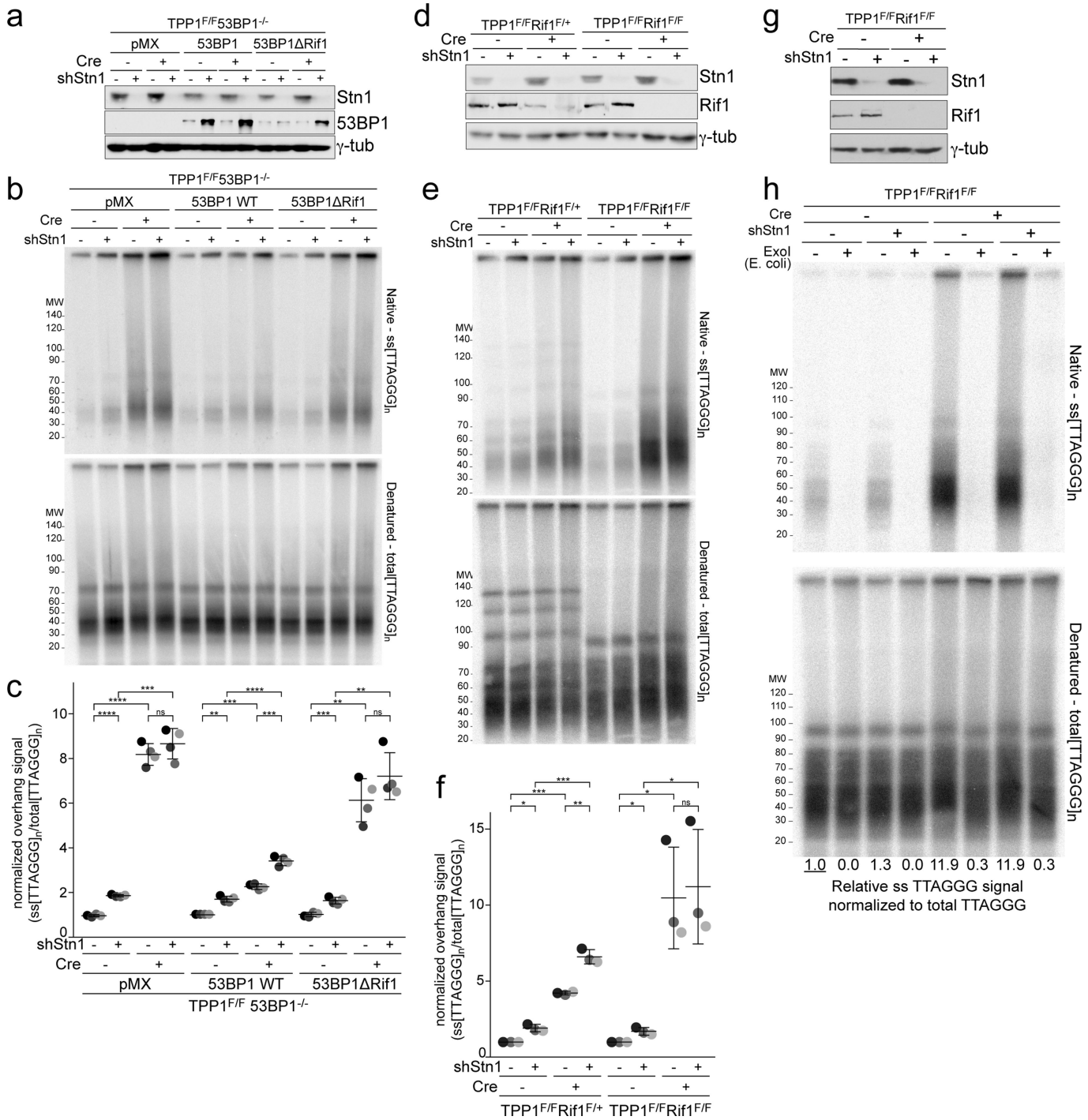
Extended Data Fig. 1 | Shieldin and CST counteract telomere hyper-resection. a-c, Effect of SHLD2 on hyper-resection at telomeres that lack TPP1. **a**, Immunoblot for CHK1-P, an indicator of TPP1 deletion, in *Tpp1*$^{f/f}$ MEFs with and without bulk population treatment with a *Shld2* sgRNA and/or Cre (representative of three experiments). **b**, Quantitative analysis of telomere end resection as in Fig. 1c using the cells shown in **a**. **c**, Quantification of the extent of resection detected in **b**, as in Fig. 1d. Mean (centre bars) and s.d. (error bars) from four independent experiments. *$P < 0.05$, **$P < 0.01$, two-tailed Welch's *t*-test. **d**, Fluorescence-activated cell sorting (FACS) profiles of the indicated cells incubated with BrdU to measure S phase effects of the *Stn1* shRNA. Gating strategy for live cells and singlets is shown below the FACS profiles. Representative of two experiments. **e**, **f**, Experiments to verify that the single-stranded DNA signal derives from a 3′ overhang. **e**, Immunoblot for STN1 and γ-tubulin in *Tpp1*$^{f/f}$ (*Rif1*$^{f/+}$) cells treated with *Stn1* shRNA and/or Cre. Representative of two experiments. **f**, Quantitative assay for telomeric overhangs, as in Fig. 1c. Plugs in the ExoI lanes were treated with the 3′ exonuclease from *E. coli*. Representative of two experiments.

© 2018 Springer Nature Limited. All rights reserved.



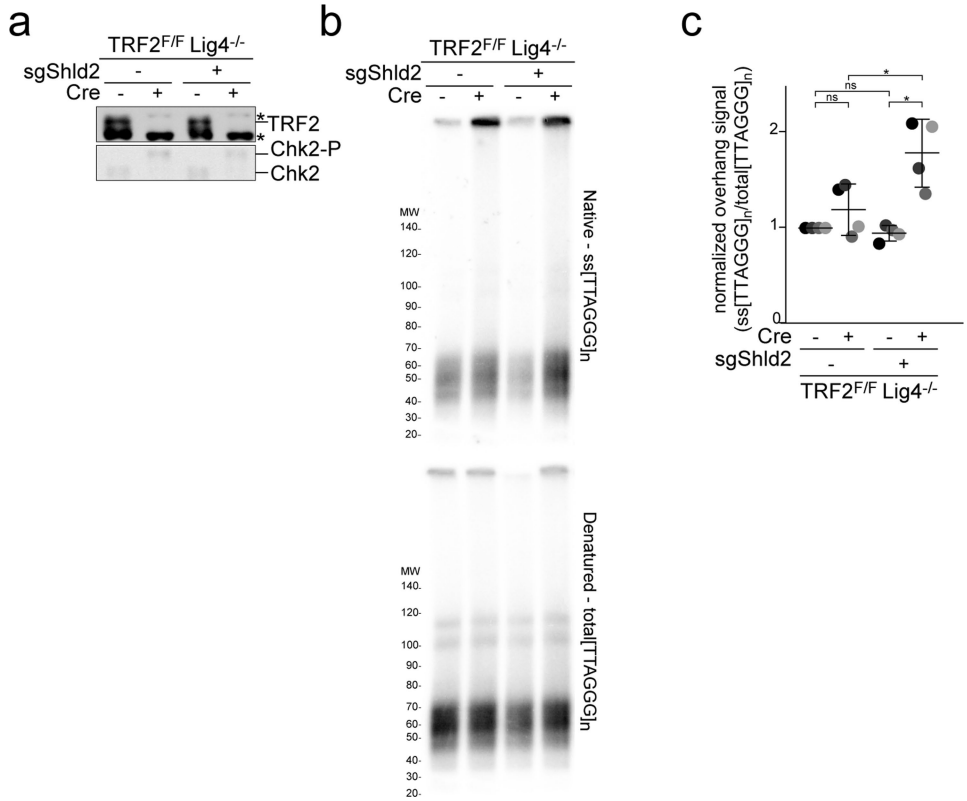
Extended Data Fig. 2 | Hyper-resection at telomeres that lack TPP1 is counteracted by CST and shieldin. a, Immunoblots showing absence of REV7 and reduction of STN1 expression in the indicated *Tpp1*$^{f/f}$ and *Tpp1*$^{f/f}$*Rev7*$^{-/-}$ MEFs treated with either *Ctc1* or *Stn1* shRNA. Diminished STN1 expression is used as a proxy for the efficacy of the *Ctc1* shRNA. Representative of two experiments. **b**, Quantitative analysis of telomeric overhangs, as in Fig. 1c. Representative of two experiments. **c**, Quantification of the effect of *Ctc1* and *Stn1* shRNA on resection at telomeres that lack TPP1, as in Fig. 1d. Data are obtained from two independent REV7-proficient and two independent REV7-deficient clones (light and dark shading).

© 2018 Springer Nature Limited. All rights reserved.



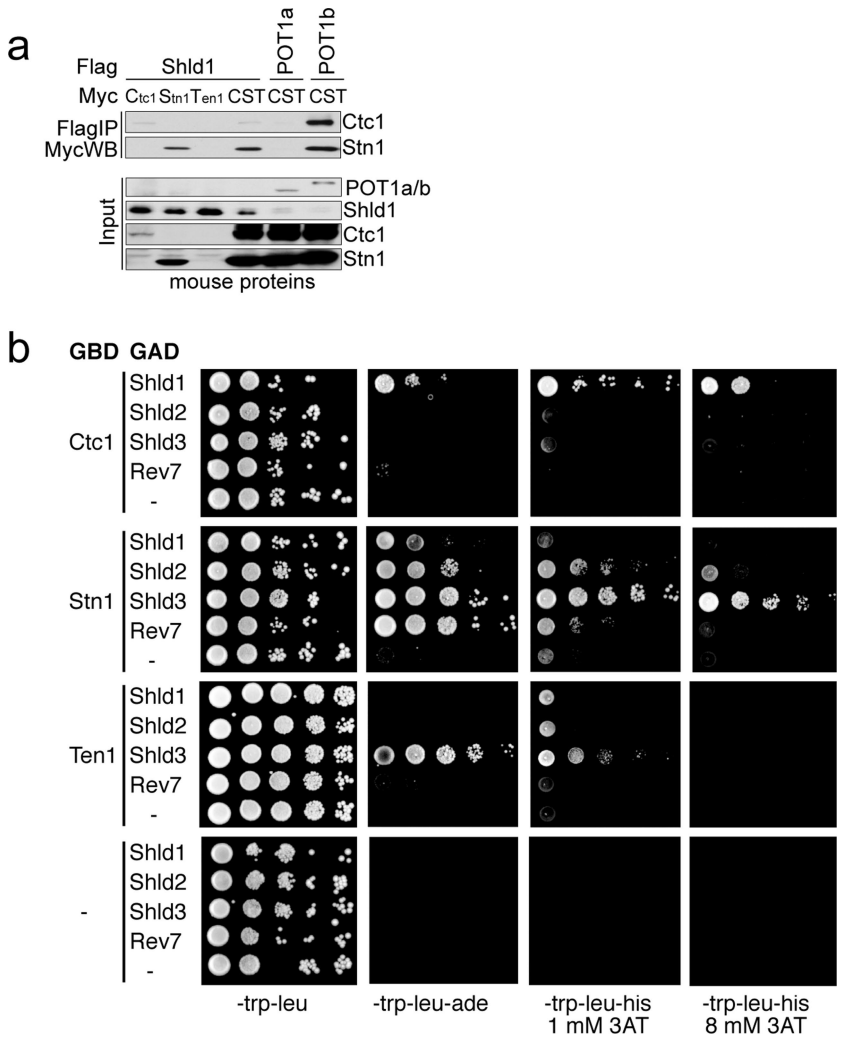
Extended Data Fig. 3 | No effect of CST depletion on telomere hyper-resection when 53BP1 or RIF1 are absent. a, SV40LT-immortalized *Tpp1*$^{f/f}$*53bp1*$^{-/-}$ cells were complemented with wild-type 53BP1 or a mutant 53BP1 that lacks the ability to interact with RIF1, treated with a *Stn1* shRNA as indicated and analysed by immunoblotting for 53BP1 and STN1. Representative of four experiments. **b**, Quantitative analysis of telomeric overhangs, as in Fig. 1c. **c**, Quantification of the resection at telomeres that lack TPP1, in four independent experiments performed as in Fig. 1d. **d**, Immunoblots showing loss of RIF1 and STN1 in the indicated *Tpp1*$^{f/f}$*Rif1*$^{f/+}$ and *Tpp1*$^{f/f}$*Rif1*$^{f/f}$ MEFs treated with Cre (96 h) as indicated, and with or without *Stn1* shRNA. Note the diminished levels of RIF1 after Cre, owing to heterozygosity in the *Tpp1*$^{f/f}$*Rif1*$^{f/+}$ cells. **e**, Quantitative analysis of telomeric overhangs, as in Fig. 1c. **f**, Quantification of the extent of resection detected, as in **e**, determined from three independent experiments (indicated by different shades of grey) showing mean (centre bars) and s.d. (error bars). Each experiment involved all indicated samples analysed in parallel. **g, h**, Experiments to verify that the single-stranded DNA signal derives from a 3′ overhang. **g**, Immunoblot for STN1 and γ-tubulin in *Tpp1*$^{f/f}$*Rif1*$^{f/f}$ cells treated with *Stn1* shRNA and/or Cre. Representative of two experiments. **h**, Quantitative assay for telomeric overhangs, as in Fig. 1c. Plugs in the ExoI lanes were treated with the 3′ exonuclease from *E. coli*. Representative of two experiments. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$, two-tailed Welch's *t*-test.

© 2018 Springer Nature Limited. All rights reserved.



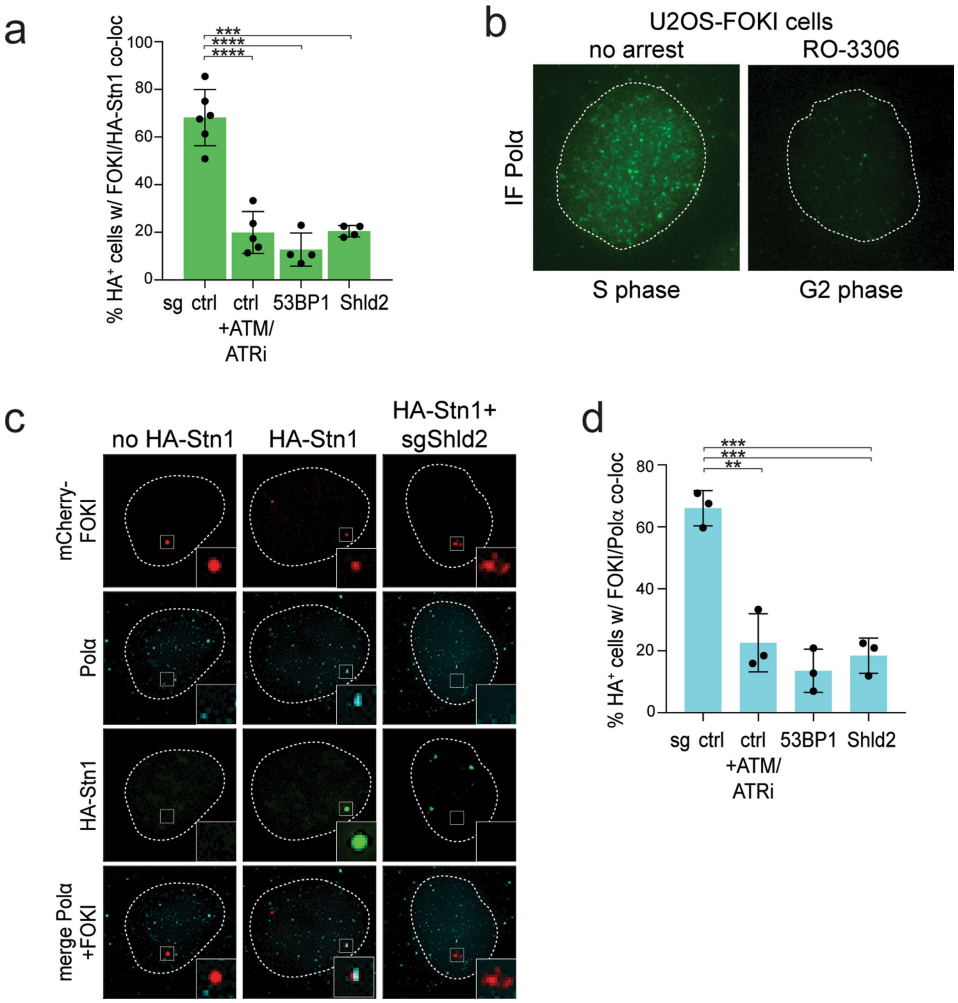
Extended Data Fig. 4 | SHLD2 counteracts resection at telomeres that lack TRF2. a, Immunoblots for TRF2 deletion and CHK2 phosphorylation in *Trf2*$^{f/f}$*Lig4*$^{-/-}$ MEFs, with and without bulk population treatment with a *Shld2* sgRNA and/or Cre. Asterisk, non-specific band. Representative of three experiments. **b**, Quantitative analysis of telomere end resection, as in Fig. 1c, using the cells shown in **a**. **c**, Quantification of the extent of resection detected in **b**, as in Fig. 1d. Mean (centre bars) and s.d. (error bars) from four independent experiments. *$P < 0.05$, two-tailed Welch's *t*-test.

© 2018 Springer Nature Limited. All rights reserved.



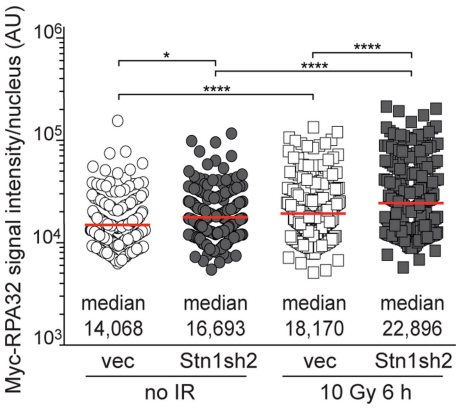
Extended Data Fig. 5 | CST interacts with shieldin. a, Immunoprecipitation of individual mouse CST subunits or the three subunit complex (each subunit bearing a Myc tag) with Flag-tagged mouse SHLD1, co-expressed in 293T cells. Flag-tagged POT1B and POT1A serve as positive and negative controls for CST binding, respectively. Representative of two experiments. **b**, Two-hybrid analysis of CST-shieldin interaction. Yeast cultures were grown overnight in synthetic complete medium that lacked tryptophan and leucine, to a density of 5×10^7 cells per millilitre. Serial tenfold dilutions were generated and 4 μl of each dilution was spotted on synthetic complete medium that lacked the nutrients tryptophan, leucine, adenine and histidine, and contained 3-aminotriazole (3-AT), as indicated. Plates were then incubated for 5 days at 30 °C before imaging. Representative of three experiments.

© 2018 Springer Nature Limited. All rights reserved.



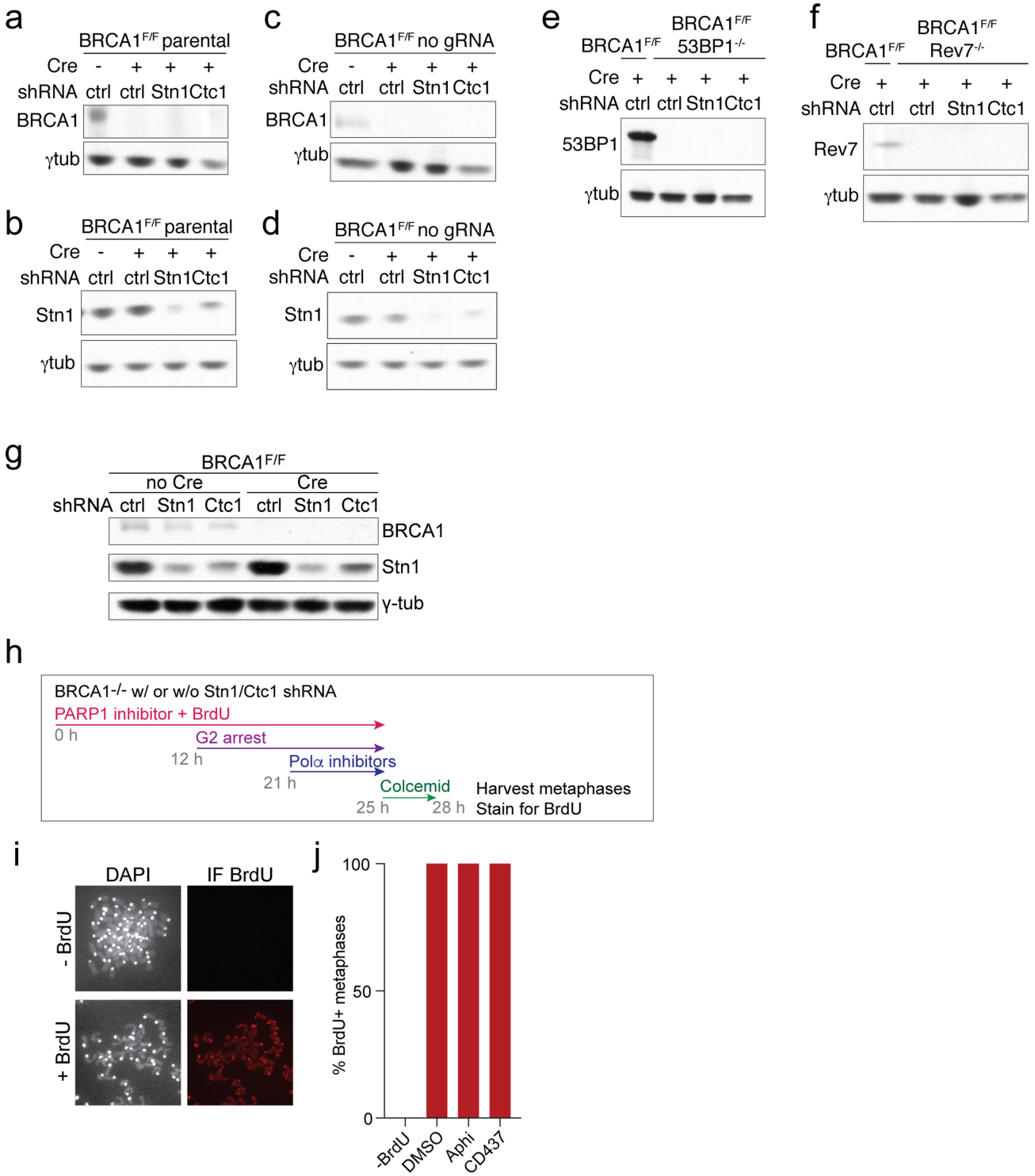
Extended Data Fig. 6 | Localization of CST and Polα to DSBs. **a**, Quantification of HA–STN1 localization to DSBs induced by FOKI, as in Fig. 3e. Mean (centre bars) and s.d. (error bars) from 4–6 independent experiments, with >80 induced nuclei for each condition in each experiment. **b**, Immunofluorescence for endogenous Polα in FOKI–LacI U2OS cells in S phase and after RO3306 treatment (G2). Dotted lines denote the outline of the nucleus. Representative of two experiments. **c**, Examples of HA–STN1 and Polα localization at DSBs induced by FOKI in G2-arrested FOKI–LacI U2OS cells (as in Fig. 3f). Representative of three experiments. **d**, Quantification of co-localization of Polα with DSBs induced by FOKI (as in Fig. 3f). Mean (centre bars) and s.d. (error bars) from three independent experiments, with >80 induced nuclei for each condition in each experiment. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$, two-tailed Welch's *t*-test.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Effect of STN1 knockdown on the intensity of RPA foci induced by ionizing radiation. Quantification of Myc–RPA32 intensity per nucleus in the experiments shown in Fig. 3g, h. Medians (centre bars and numbers below) obtained from four independent experiments, with >20 nuclei for each experimental condition in each experiment. Each symbol represents one nucleus. *$P < 0.05$, ****$P < 0.0001$, two-tailed Welch's t-test.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Effect of CST and Polα on PARPi treatment of BRCA1-deficient cells. a–f, Immunoblots on the MEFs used in Fig. 4a–e to verify the absence of deleted proteins and efficacy of the shRNAs. Reduction in STN1 expression is used as a proxy for the efficacy of the *Ctc1* shRNA because no antibody to mouse CTC1 is available. Each immunoblot is representative of three experiments. **g**, Immunoblots for BRCA1 and STN1 in the cells used in Fig. 4f. Representative of two experiments. **h–j**, Control experiment to assess that cells analysed in Fig. 4f progressed through S phase during treatment with PARPi. **h**, Experimental timeline, as in Fig. 4f, but with inclusion of BrdU in the medium during treatment with PARPi. **i**, Example of the assay for the presence of BrdU (immunofluorescence) in metaphases collected after the experimental timeline, as in **h**. **j**, Quantification of the BrdU incorporation into metaphase chromosomes, as in **i** (one experiment with ten metaphases per condition).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0340-7

The shieldin complex mediates 53BP1–dependent DNA repair

Sylvie M. Noordermeer[1,2,10], Salomé Adam[1,10], Dheva Setiaputra[1,10], Marco Barazas[3], Stephen J. Pettitt[4], Alexanda K. Ling[5], Michele Olivieri[1,6], Alejandro Álvarez-Quilón[1], Nathalie Moatti[1], Michal Zimmermann[1], Stefano Annunziato[3], Dragomir B. Krastev[4], Feifei Song[4], Inger Brandsma[4], Jessica Frankum[4], Rachel Brough[4], Alana Sherker[1,6], Sébastien Landry[1], Rachel K. Szilard[1], Meagan M. Munro[1], Andrea McEwan[1], Théo Goullet de Rugy[1], Zhen-Yuan Lin[1], Traver Hart[7], Jason Moffat[6,8], Anne-Claude Gingras[1,6], Alberto Martin[5], Haico van Attikum[2], Jos Jonkers[3], Christopher J. Lord[4], Sven Rottenberg[3,9] & Daniel Durocher[1,6]*

53BP1 is a chromatin-binding protein that regulates the repair of DNA double-strand breaks by suppressing the nucleolytic resection of DNA termini[1,2]. This function of 53BP1 requires interactions with PTIP[3] and RIF1[4–9], the latter of which recruits REV7 (also known as MAD2L2) to break sites[10,11]. How 53BP1-pathway proteins shield DNA ends is currently unknown, but there are two models that provide the best potential explanation of their action. In one model the 53BP1 complex strengthens the nucleosomal barrier to end-resection nucleases[12,13], and in the other 53BP1 recruits effector proteins with end-protection activity. Here we identify a 53BP1 effector complex, shieldin, that includes C20orf196 (also known as SHLD1), FAM35A (SHLD2), CTC-534A2.2 (SHLD3) and REV7. Shieldin localizes to double-strand-break sites in a 53BP1- and RIF1-dependent manner, and its SHLD2 subunit binds to single-stranded DNA via OB-fold domains that are analogous to those of RPA1 and POT1. Loss of shieldin impairs non-homologous end-joining, leads to defective immunoglobulin class switching and causes hyper-resection. Mutations in genes that encode shieldin subunits also cause resistance to poly(ADP-ribose) polymerase inhibition in BRCA1-deficient cells and tumours, owing to restoration of homologous recombination. Finally, we show that binding of single-stranded DNA by SHLD2 is critical for shieldin function, consistent with a model in which shieldin protects DNA ends to mediate 53BP1-dependent DNA repair.

To discover proteins acting in the 53BP1 pathway, we searched for genes whose mutation restores homologous recombination in BRCA1-deficient cells and leads to resistance to poly(ADP-ribose) polymerase (PARP) inhibition, which is a hallmark of 53BP1 deficiency[14–16]. We undertook three independent CRISPR–Cas9 screens that entailed the transduction of BRCA1-deficient cells with lentiviral libraries of single-guide RNAs (sgRNAs) (Extended Data Fig. 1a). The resulting pools of edited cells were exposed to near-lethal doses of two clinically used PARP inhibitors (PARPi), either olaparib or talazoparib[17]. We screened both an engineered human RPE1-hTERT *TP53*$^{-/-}$*BRCA1*$^{-/-}$ cell line (hereafter, RPE1 *BRCA1*^{KO}) and SUM149PT cells carrying a hemizygous *BRCA1* frameshift mutation. The gene-based results of the olaparib screens are found in Supplementary Table 1 and those of the talazoparib screen were published previously[18].

The genes coding for 53BP1 and for the uncharacterized protein C20orf196 were hits in all three screens (Fig. 1a). We also identified *SCAF1* and *ATMIN*, which encode an SR-family protein and a transcription factor, as hits in the two olaparib screens (Fig. 1a) whereas genes coding for proteins acting upstream (H2AX, MDC1, RNF8 and RNF168) or downstream (RIF1) of 53BP1 were hits in the RPE1 *BRCA1*^{KO} screen (Supplementary Table 1). The presence of 53BP1 and 53BP1-pathway proteins suggested that these screens could reveal hitherto uncharacterized 53BP1 effectors.

In competitive growth assays (Fig. 1b), sgRNAs targeting *53BP1* (also known as *TP53BP1*) led to the outgrowth of *BRCA1*^{KO} cells in the presence of olaparib (Fig. 1c; genotyping information in Supplementary Table 2). Similarly, sgRNAs targeting *C20orf196*, *ATMIN* and *SCAF1* led to resistance to PARPi (Fig. 1c and Extended Data Fig. 1b). In parallel studies, transfection of CRISPR RNAs (crRNAs) and *trans*-activating crRNAs (tracrRNAs) that target *C20orf196*, *53BP1* or *PARP1* caused talazoparib resistance in SUM149PT cells (Fig. 1d and Extended Data Fig. 1c). Because C20orf196 was identified as a hit in all three screens and validated in independent assays, we sought to determine its role in DNA repair.

C20orf196 is an uncharacterized protein comprising 205 amino acid residues (Fig. 1e), previously identified as a candidate REV7-interacting protein[19]. We used immunoprecipitation coupled to mass spectrometry (IP-MS) to expand the interaction network surrounding the REV7-interacting proteins (Fig. 1f and Supplementary Table 3). One protein, FAM35A, was enriched in both C20orf196 and REV7 immunoprecipitation samples (Fig. 1f). FAM35A was notable owing to the presence of three predicted OB-fold domains (OBA, OBB and OBC; Fig. 1e), reminiscent of those in the single-stranded DNA (ssDNA)-binding proteins RPA1[20] and POT1[21]. IP-MS experiments with FAM35A recovered CTC-534A2.2, also identified in the REV7 IP-MS (Fig. 1f and Supplementary Table 3). CTC-534A2.2 is a protein encoded by an alternative mRNA emanating from the *TRAPPC13* locus (Fig. 1e and Extended Data Fig. 1d). sgRNAs against *CTC-534A2.2* were not present in either of our first-generation sgRNA libraries. IP-MS with CTC-534A2.2 recovered C20orf196, FAM35A and REV7 (Fig. 1f and Supplementary Table 3), which suggests that these proteins form a single protein complex—this was confirmed by sequential affinity purification of epitope-tagged C20orf196, CTC-534A2.2, FAM35A and REV7 (Fig. 1g).

FAM35A, *C20orf196* and *CTC-534A2.2* were identified in a fourth CRISPR–Cas9 screen using a second-generation sgRNA library, TKOv2. This screen sought to identify genes that promote resistance to ionizing radiation in RPE1 cells (Extended Data Fig. 1e). Seventy-five genes scored at a false discovery rate (FDR) $< 1\%$ and this gene set was found to be highly enriched in genes that encode non-homologous end-joining (NHEJ) factors by Gene Ontology enrichment ($P = 1.11 \times 10^{-11}$, Fisher's exact test with multiple correction;

[1]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [2]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [3]Division of Molecular Pathology, Oncode Institute, Netherlands Cancer Institute, Amsterdam, The Netherlands. [4]The CRUK Gene Function Laboratory and Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK. [5]Department of Immunology, University of Toronto, Toronto, Ontario, Canada. [6]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [7]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [8]Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [9]Institute of Animal Pathology, Vetsuisse Faculty, University of Bern, Bern, Switzerland. [10]These authors contributed equally: Sylvie M. Noordermeer, Salomé Adam, Dheva Setiaputra. *e-mail: durocher@lunenfeld.ca

© 2018 Springer Nature Limited. All rights reserved.

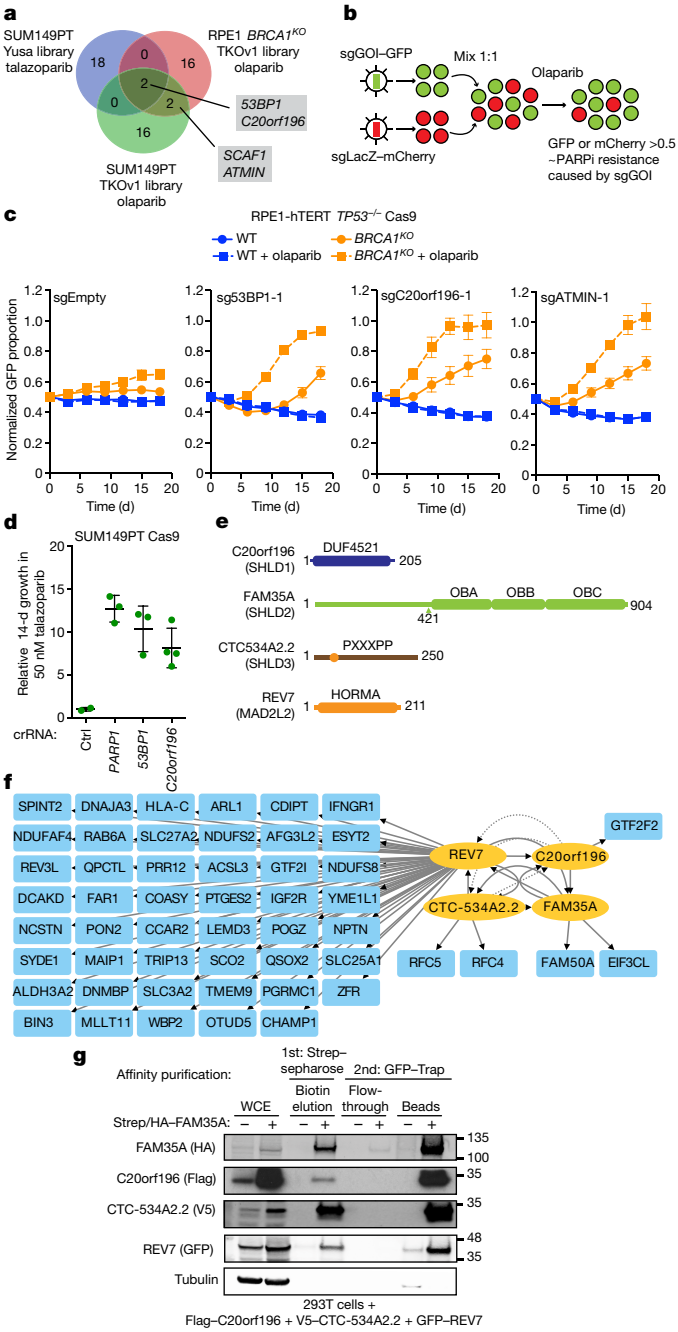


Fig. 1 | Identification of shieldin. a, Venn diagram of the top 20 hits in each screen. **b**, Schematic of the competitive growth assays. **c**, Competitive growth assays ± olaparib (16 nM) in RPE1 *BRCA1*^{KO} cells. Data represent mean fraction of GFP-positive cells ± s.e.m., normalized to day 0 ($n = 3$, independent transductions). **d**, Resistance to PARPi caused by mutation of *C20orf196*. SUM149PT cells transfected with indicated crRNAs were treated with 50 nM talazoparib for 14 d. Relative growth was normalized to a non-targeting crRNA (Ctrl). Bars represent mean ± s.d. of multiple crRNAs per gene (shown is a representative plot of 2 biologically independent experiments). **e**, Domain architecture of shieldin subunits. SHLD3 contains a REV7-binding PXXXPP motif[28]. **f**, Protein interaction network surrounding REV7, C20orf196, FAM35A and CTC-534A2.2. Solid and dashed arrows represent interactions at an FDR of ≤ 1% and ≤ 5%, respectively. **g**, Sequential affinity purifications from 293T cell lysates expressing the indicated proteins. Bound proteins were immunoblotted with the indicated antibodies ($n = 2$ independent experiments). WCE, whole-cell extract.

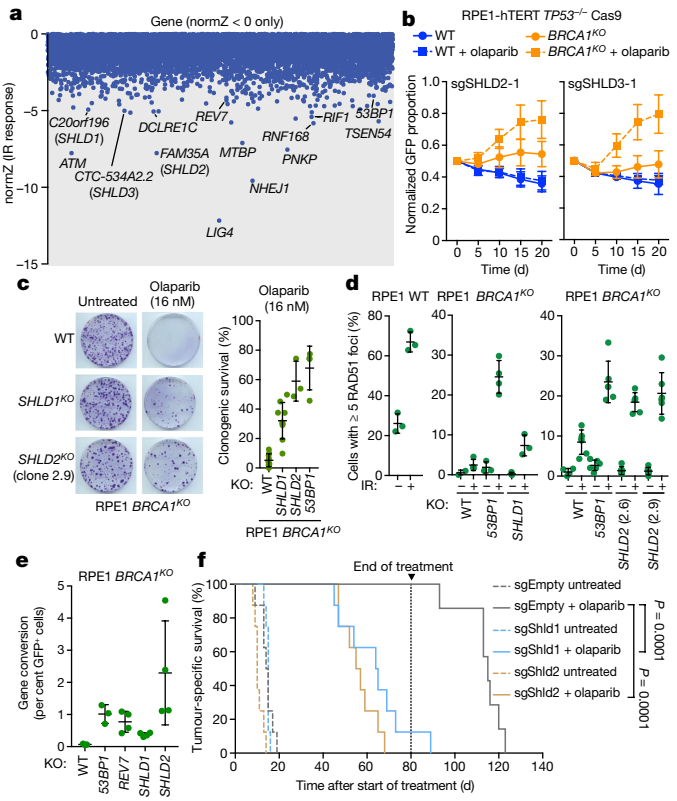


Fig. 2 | Shieldin loss promotes PARPi resistance in cell and tumour models of *BRCA1*-deficiency. a, CRISPR dropout screen results in RPE1 wild-type cells exposed to ionizing radiation. Gene-level normZ scores < 0 are shown. **b**, Competitive growth assays using olaparib (16 nM) in RPE1 *BRCA1*^{KO} cells. Data are presented as mean ± s.d., normalized to day 0 ($n = 3$, independent transductions). **c**, Clonogenic survival in response to 16 nM olaparib. Representative images are shown (left) and quantified (right). Bars represent mean ± s.d. ($n = 9$: RPE1 wild type and *BRCA1*^{KO}*SHLD1*^{KO}; $n = 3$: *BRCA1*^{KO}*SHLD2*^{KO}, $n = 4$: *BRCA1*^{KO}*53BP1*^{KO}; biologically independent experiments). **d**, Quantification of cells with ≥ 5 RAD51 foci ± 10 Gy ionizing radiation (6 h recovery). Biologically independent experiments are shown and the bars represent the mean ± s.d. From left to right, the number of replicates was $n = 3$ (for both conditions in left panel); $n = 3, 4, 3, 4, 3$ and 3 (for conditions in middle panel, left to right); and $n = 4, 6, 6, 6, 6, 6, 6$ and 6 (for conditions in right panel, left to right). **e**, Assessment of gene conversion by traffic light reporter assay. Biologically independent experiments are shown and the bars represent the mean ± s.d. ($n = 3$ for wild type and *53BP1*^{KO}; $n = 4$ for *SHLD1*^{KO}, *SHLD2*^{KO} and *REV7*^{KO}). **f**, Kaplan–Meier curve showing tumour-specific survival of mice transplanted with KB1P4 tumour organoids ± olaparib treatment for 80 d ($n = 8$ per treatment; editing efficiencies found in Supplementary Table 2). *P* values were calculated using a log-rank test (Mantel–Cox). IR, ionizing radiation; KO, knock out; WT, wild type.

Fig. 2a and Supplementary Table 4). *RIF1*, *FAM35A*, *C20orf196*, *CTC-534A2.2*, *53BP1* and *REV7* were all hits at FDR < 1% (Fig. 2a). These data suggest that the complex formed by C20orf196, FAM35A, REV7 and CTC-534A2.2 promotes the repair of double-strand breaks (DSBs) by NHEJ. For reasons that will become apparent, we named this complex shieldin and renamed *C20orf196*, *FAM35A* and *CTC-534A2.2* as *SHLD1*, *SHLD2* and *SHLD3*, respectively.

Independent sgRNAs targeting *SHLD2* or *SHLD3* caused sensitivity to the clastogen etoposide in competitive growth assays (Extended Data Fig. 1f) and caused resistance to olaparib in RPE1 *BRCA1*^{KO} cells, consistent with SHLD2 and SHLD3 acting with REV7 and SHLD1 (Fig. 2b and Extended Data Fig. 1g). Clonal knockouts of *SHLD1* or *SHLD2* led to olaparib resistance in *BRCA1*^{KO} cells, and *SHLD2*^{KO} resulted in a phenotype that approached that of 53BP1 loss (Fig. 2c). Similar results were obtained with 11 independent clonal knockouts of *SHLD1* in SUM149PT cells exposed to talazoparib (Extended Data Fig. 1h). Furthermore, expressing GFP–SHLD2 in double knockout *BRCA1*^{KO}*SHLD2*^{KO} cells restored olaparib sensitivity (Extended Data

© 2018 Springer Nature Limited. All rights reserved.

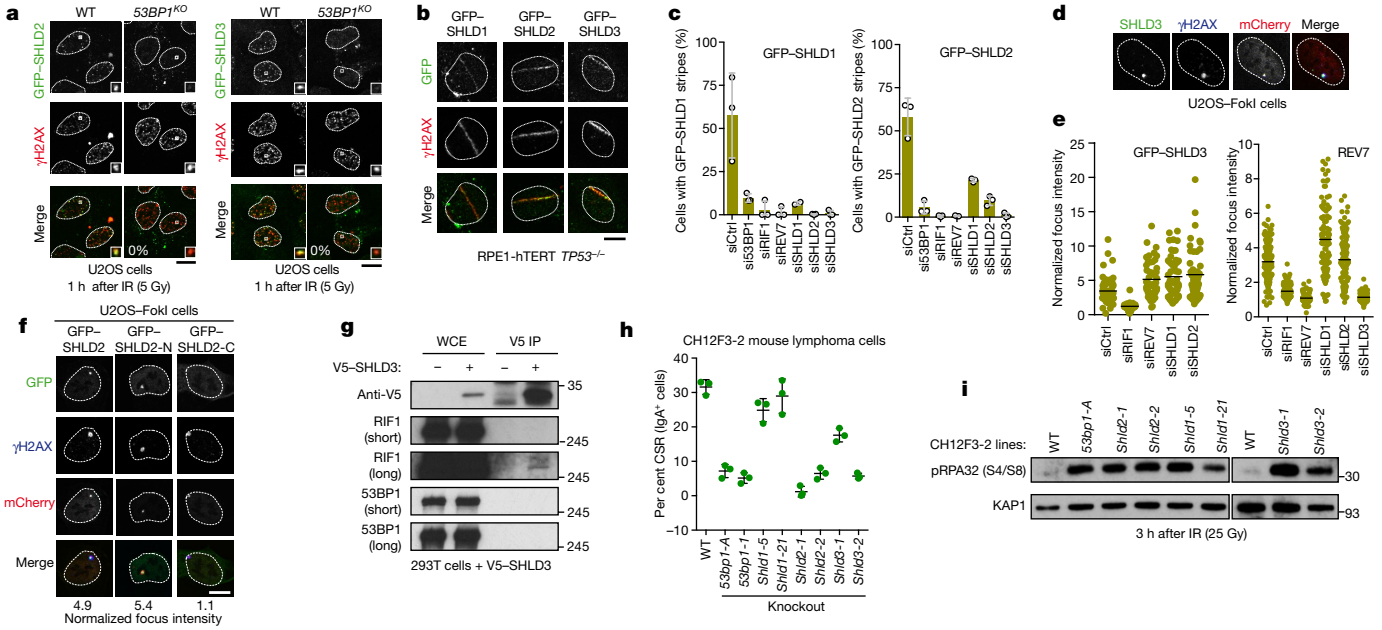


Fig. 3 | Shieldin accumulates at DSB sites downstream of 53BP1–RIF1 and promotes class switch recombination. a, **b**, Representative micrographs of shieldin subunit accumulation into foci induced by ionizing radiation in U2OS cells (**a**), or at laser microirradiation sites in RPE1 cells (**b**) ($n = 3$ biologically independent experiments). **c**, Quantification of **b**. Points represent individual experiments counting ≥100 cells. Data are mean ± s.d. **d**, Colocalization of GFP-tagged shieldin subunits with mCherry foci in U2OS–FokI cells upon mCherry–LacR–FokI expression. **e**, Quantification of GFP–SHLD3 or endogenous REV7 focus intensity. Each point represents a cell. Lines represent the mean. Data are the aggregate of two biological replicates with a total of 50 and 120 (GFP–SHLD3 and REV7, respectively), 45 and 122, 48 and 112, 54 and 116, 49 and 111, and 117 (for *SHLD3* siRNA, only REV7 foci were quantified) cells counted for control siRNA and siRNAs against *RIF1* (siRIF1), *REV7* (siREV7), *SHLD1* (siSHLD1), *SHLD2* (siSHLD2) and *SHLD3* (siSHLD3), respectively. **f**, mCherry–LacR–FokI colocalization with GFP–SHLD2, the N terminus of SHLD2 (SHLD2-N), and SHLD2-C. Mean normalized focus intensity is shown with a total of 52 (SHLD2 and SHLD2-N) or 53 (SHLD2-C) cells counted ($n = 2$, biologically independent experiments). **g**, RIF1 co-immunoprecipitation with V5–SHLD3 (representative of two independent experiments). **h**, Class switch recombination analysis of CH12F3-2 cells following stimulation. Data are mean ± s.d. ($n = 3$ biologically independent replicates, clone number of the knockout cell lines is indicated by the number following the gene name). **i**, Irradiated (25 Gy) CH12F3-2 clones were immunoblotted for RPA32 phosphorylation (pRPA32, phosphorylated at Ser4 and Ser8) (representative of three biological replicates, clone number of the knockout cell lines is indicated by the number following the gene name).

Fig. 1i). Resistance to PARPi in shieldin-deficient *BRCA1*^{KO} cells is due to restoration of homologous recombination, as measured both by RAD51 focus formation induced by ionizing radiation and by a reporter for gene conversion (the traffic light reporter assay[22]) (Fig. 2d, e and Extended Data Fig. 2a–d).

Next, we tested whether loss of shieldin causes resistance to PARPi in the KB1P mouse mammary tumour model that is deficient in *Brca1* and *Trp53*, which encodes p53 (ref. [23]). sgRNAs targeting *Shld1* and *Shld2* led to resistance to PARPi in KB1P-G3 mammary tumour cells and in *Brca1*$^{-/-}$*Trp53*$^{-/-}$ mouse embryonic stem cells (Extended Data Fig. 3a, c). This resistance was also associated with restoration of homologous recombination (Extended Data Fig. 3b). Furthermore, transduction of *Shld1*- and *Shld2*-targeting sgRNAs suppressed the cell lethality associated with *Brca1* loss in p53-proficient mouse embryonic stem cells (Extended Data Fig. 3d). We transduced the same sgRNAs into KB1P4 mammary tumour organoids (Supplementary Table 2) and implanted them into the fat pads of mice. Olaparib treatment was initiated when tumours reached sizes of 50–100 mm^3, and was continued for 80 days. Although all untreated mice succumbed to excessive tumour burden within 20 days, the control group responded to olaparib for the duration of the treatment (Fig. 2f). However, mice implanted with *Shld1*- and *Shld2*-mutated tumours exhibited a partial response to olaparib, with mice succumbing by day 60 (Fig. 2f). We conclude that shieldin loss causes resistance to PARPi in both human and mouse BRCA1-deficient tumour cells by reactivating homologous recombination.

As expected of a complex with a direct role in DSB repair, shieldin accumulates at DSB sites in a 53BP1- and RIF1-dependent manner (Fig. 3a–f and Extended Data Fig. 4a–d). Loss of shieldin components did not impair formation of 53BP1 or RIF1 foci induced by ionizing radiation, indicating that shieldin acts downstream of 53BP1–RIF1 (Extended Data Fig. 4e, f). Consistent with this possibility, we observed genetic epistasis between *53BP1* and the shieldin genes using the RAD51 focus-formation assay in RPE1 *BRCA1*^{KO} cells (Extended Data Fig. 5a). We also observed that *SHLD1* and *53BP1* were epistatic in terms of modulating talazoparib resistance in SUM149PT cells (Extended Data Fig. 5b).

Analyses of the dependencies within the shieldin complex indicate that SHLD3 is the most apical component followed by REV7, SHLD2 and then SHLD1 (Fig. 3c–e, Extended Data Figs. 4a–d, 6, 7c–e and Supplementary Note 1 for mapping details). Indeed, SHLD3 interacts with RIF1, which suggests that SHLD3 recruits shieldin to chromatin-bound 53BP1–RIF1 (Fig. 3g and Extended Data Fig. 7f). Further mapping studies suggest that shieldin consists of a DSB-recruitment module composed of SHLD3–REV7 that binds to the N terminus of SHLD2 (residues 1–50; Extended Data Figs. 6, 7a–c), and a presumptive DNA-binding module (SHLD2–SHLD1) that features the OB-fold domains at the SHLD2 C terminus (hereafter SHLD2-C, residues 421–904; Extended Data Fig. 7a).

To assess the role of shieldin in NHEJ, we first analysed class switch recombination in CH12F3-2 cells[24]. Mutation of each of the shieldin subunits compromised class switch recombination, with *Shld1*-edited cells having a reproducibly milder phenotype (Fig. 3h and Extended Data Fig. 8a–c). *Shld2*^{KO} was epistatic with both *53bp1*^{KO} and *Shld1*^{KO} mutations, consistent with them acting in the same genetic pathway (Extended Data Fig. 8b, c). The expression of AID—which initiates class switch recombination—was not altered in shieldin mutants, consistent with NHEJ deficiency (Extended Data Fig. 8d). Supporting this possibility, *SHLD1* and *SHLD2* mutations impaired random plasmid integration—which occurs largely through NHEJ[25]—to an extent similar to that of 53BP1-deficient cells (Extended Data Fig. 8e, f).

© 2018 Springer Nature Limited. All rights reserved.

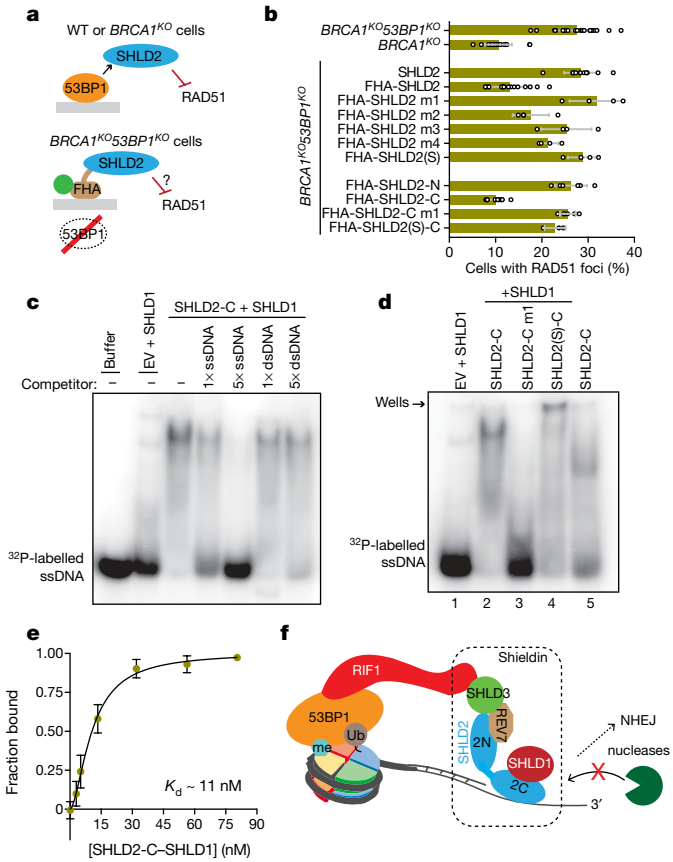


Fig. 4 | Shieldin is an effector of 53BP1 by binding ssDNA. **a**, Schematic for artificially targeting shieldin to DSB sites. **b**, Formation of RAD51 foci induced by ionizing radiation 3 h after 10 Gy irradiation in *BRCA1*^{KO}*53BP1*^{KO} cells expressing the indicated fusion proteins. Data are mean ± s.d. From top to bottom, the number of biologically independent experiments was $n = 20$, 22, 12, 12, 16, 4, 4, 4, 4, 4, 6, 6 and 3. **c**, EMSA of the SHLD2-C–SHLD1 complex isolated from 293T cells (see Extended Data Fig. 9b) incubated with radio-labelled ssDNA ± unlabelled oligonucleotides ($n = 2$ independent experiments). EV, empty vector. **d**, EMSA of SHLD2-C wild type and variants ($n = 4$ independent experiments). **e**, Determination of SHLD2-C–SHLD1 ssDNA binding dissociation constant (K_d). Data are mean ± s.d. ($n = 3$ independent experiments). Representative EMSA shown in Extended Data Fig. 9c. **f**, Model of shieldin function. We speculate that the SHLD2 OB-fold domains bind to ssDNA at DSB sites to suppress resection and favour NHEJ.

The loss of each shieldin subunit led to ionizing-radiation-induced RPA32 Ser4/Ser8 phosphorylation, which is a surrogate marker of end-resection[20], suggesting that shieldin protects DNA ends (Fig. 3i and Extended Data Fig. 8g). Supporting this hypothesis, the restoration of homologous recombination in shieldin-defective KB1P-G3 cells was dependent on ATM activity (Extended Data Fig. 3a, b), which promotes DNA end-resection in the absence of 53BP1[16] or REV7[10]. An accompanying paper[26] shows that *Shld2*-mutated cells have increased end-resection at dysfunctional telomeres. Shieldin therefore antagonizes end-resection.

We surmised that if shieldin is a downstream effector of 53BP1, artificially targeting shieldin to DSB sites should rescue phenotypes associated with 53BP1 loss. To do this, we fused SHLD2 to the RNF8 forkhead-associated (FHA) domain, which is recruited to damaged chromatin independently of 53BP1 (Fig. 4a). We found that the FHA-dependent targeting of SHLD2 to DSB sites suppressed RAD51 focus formation induced by ionizing radiation in *BRCA1*^{KO}*53BP1*^{KO} cells (Fig. 4b and Extended Data Fig. 9a). These results suggest that SHLD2 mediates 53BP1-dependent DNA repair.

We observed that the FHA–SHLD2-C protein, which contains the OB-fold domains, potently suppressed RAD51 recruitment to DSB sites in *BRCA1*^{KO}*53BP1*^{KO} cells (Fig. 4b and Extended Data Fig. 9a). This result suggested that DNA binding might underpin the effector function of the SHLD2 C terminus. To test for DNA binding activity, we affinity-purified SHLD2-C in the presence or absence of SHLD1 (Extended Data Fig. 9b). We observed SHLD2-C binding to a radio-labelled ssDNA probe by electrophoretic mobility shift assays (EMSA), and competition with unlabelled oligonucleotides revealed that SHLD2-C preferentially binds to ssDNA over double-stranded DNA (Fig. 4c). Although SHLD1 is not essential for the DNA-binding activity of SHLD2-C, its presence increased the amount of SHLD2-C purified, and the retarded complex displayed a difference in mobility consistent with the SHLD2-C–SHLD1 complex binding to ssDNA (Fig. 4d, lane 2 versus 5). We estimate the binding affinity of the SHLD2-C–ssDNA interaction to be about 10 nM (Fig. 4e and Extended Data Fig. 9c). We conclude that SHLD2 possesses ssDNA-binding activity.

To explore whether ssDNA binding is involved in shieldin function, we generated four mutant versions (named m1–m4) of the SHLD2 OB folds by modelling the SHLD2 C terminus on an RPA1 structure[27] (RCSB Protein Data Bank code: 4GNX; Extended Data Fig. 9d). We also employed a splice variant of SHLD2 that disrupts OBB, which we refer to as SHLD2(S). We found that the SHLD2 m1 and SHLD2(S) mutants—either in the context of full-length SHLD2 or SHLD2-C proteins—were unable to suppress RAD51 focus formation in *BRCA1*^{KO}*53BP1*^{KO} cells when fused to the RNF8 FHA domain (Fig. 4b and Extended Data Fig. 9a). Expression of full-length SHLD2 m1 and SHLD2(S) in *BRCA1*^{KO}*SHLD2*^{KO} cells also failed to suppress RAD51 focus formation induced by ionizing radiation, unlike wild-type SHLD2 (Extended Data Fig. 10a, b). Importantly, both mutants localized to DSB sites (Extended Data Fig. 10c, d) and interacted with the other members of the shieldin complex (Extended Data Fig. 10e). Therefore, the SHLD2 m1 and SHLD2(S) mutants are defective in suppressing homologous recombination.

Notably, the SHLD2-C m1 mutant was completely defective in ssDNA binding (Fig. 4d, lane 3) whereas the SHLD2(S)-C mutant displayed reduced and aberrant ssDNA-binding behaviour (Fig. 4d, lane 4). Because the m1 mutation produces a protein that is defective both in ssDNA-binding and suppression of homologous recombination, but which is proficient in both complex assembly and DSB recruitment, we conclude that ssDNA binding by shieldin is critical for 53BP1-dependent DSB repair.

In conclusion, the identification of shieldin forces us to re-evaluate how DNA end stability is regulated in vertebrates. Our results are consistent with a model in which shieldin is the ultimate mediator of 53BP1-dependent DNA repair by binding ssDNA and occluding access to resection nucleases (Fig. 4f). Our discovery of shieldin also has implications for the management of *BRCA1*-mutated malignancies, as alterations in shieldin-coding genes may cause clinical resistance to PARPi.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0340-7

Received: 4 January 2018; Accepted: 15 May 2018;
Published online 18 July 2018.

1. Panier, S. & Boulton, S. J. Double-strand break repair: 53BP1 comes into focus. *Nat. Rev. Mol. Cell Biol.* **15**, 7–18 (2014).
2. Hustedt, N. & Durocher, D. The control of DNA repair by the cell cycle. *Nat. Cell Biol.* **19**, 1–9 (2017).
3. Munoz, I. M., Jowsey, P. A., Toth, R. & Rouse, J. Phospho-epitope binding by the BRCT domains of hPTIP controls multiple aspects of the cellular response to DNA damage. *Nucleic Acids Res.* **35**, 5312–5322 (2007).
4. Zimmermann, M., Lottersberger, F., Buonomo, S. B., Sfeir, A. & de Lange, T. 53BP1 regulates DSB repair using Rif1 to control 5′ end resection. *Science* **339**, 700–704 (2013).

© 2018 Springer Nature Limited. All rights reserved.

5. Feng, L., Fong, K. W., Wang, J., Wang, W. & Chen, J. RIF1 counteracts BRCA1-mediated end resection during DNA repair. *J. Biol. Chem*. **288**, 11135–11143 (2013).
6. Escribano-Díaz, C. et al. A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. *Mol. Cell* **49**, 872–883 (2013).
7. Di Virgilio, M. et al. Rif1 prevents resection of DNA breaks and promotes immunoglobulin class switching. *Science* **339**, 711–715 (2013).
8. Chapman, J. R. et al. RIF1 is essential for 53BP1-dependent nonhomologous end joining and suppression of DNA double-strand break resection. *Mol. Cell*, **49**, 858–871 (2013).
9. Callen, E. et al. 53BP1 mediates productive and mutagenic DNA repair through distinct phosphoprotein interactions. *Cell* **153**, 1266–1280 (2013).
10. Xu, G. et al. REV7 counteracts DNA double-strand break resection and affects PARP inhibition. *Nature* **521**, 541–544 (2015).
11. Boersma, V. et al. MAD2L2 controls DNA repair at telomeres and DNA breaks by inhibiting 5′ end resection. *Nature* **521**, 537–540 (2015).
12. Fradet-Turcotte, A. et al. 53BP1 is a reader of the DNA-damage-induced H2A Lys15 ubiquitin mark. *Nature* **499**, 50–54 (2013).
13. Adkins, N. L., Niu, H., Sung, P. & Peterson, C. L. Nucleosome dynamics regulates DNA processing. *Nat. Struct. Mol. Biol*. **20**, 836–842 (2013).
14. Jaspers, J. E. et al. Loss of 53BP1 causes PARP inhibitor resistance in *Brca1*-mutated mouse mammary tumors. *Cancer Discov*. **3**, 68–81 (2013).
15. Bouwman, P. et al. 53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers. *Nat. Struct. Mol. Biol*. **17**, 688–695 (2010).
16. Bunting, S. F. et al. 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. *Cell* **141**, 243–254 (2010).
17. Lord, C. J. & Ashworth, A. PARP inhibitors: synthetic lethality in the clinic. *Science* **355**, 1152–1158 (2017).
18. Barazas, M. et al. The CST complex mediates end protection at double-strand breaks and promotes PARP inhibitor sensitivity in BRCA1-deficient cells. *Cell Rep*. **23**, 2107–2118 (2018).
19. Hutchins, J. R. et al. Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science* **328**, 593–599 (2010).
20. Marechal, A. & Zou, L. RPA-coated single-stranded DNA as a platform for post-translational modifications in the DNA damage response. *Cell Res*. **25**, 9–23 (2014).
21. Baumann, P. & Price, C. Pot1 and telomere maintenance. *FEBS Lett*. **584**, 3779–3784 (2010).
22. Kuhar, R. et al. Novel fluorescent genome editing reporters for monitoring DNA repair pathway utilization at endonuclease-induced breaks. *Nucleic Acids Res*. **42**, e4 (2014).
23. Liu, X. et al. Somatic loss of BRCA1 and p53 in mice induces mammary tumors with features of human BRCA1-mutated basal-like breast cancer. *Proc. Natl Acad. Sci. USA* **104**, 12111–12116 (2007).
24. Nakamura, M. et al. High frequency class switching of an IgM^+ B lymphoma clone CH12F3 to IgA^+ cells. *Int. Immunol*. **8**, 193–201 (1996).
25. Luijsterburg, M. S. et al. PARP1 links CHD2-mediated chromatin expansion and H3.3 deposition to DNA repair by non-homologous end-joining. *Mol. Cell* **61**, 547–562 (2016).
26. Mirman, Z. et al. 53BP1–RIF1–shieldin counteracts DSB resection through CST- and Polα-dependent fill-in. *Nature* https://doi.org/10.1038/s41586-018-0324-7 (2018).
27. Fan, J. & Pavletich, N. P. Structure and conformational change of a replication protein A heterotrimer bound to ssDNA. *Genes Dev*. **26**, 2337–2347 (2012).
28. Hanafusa, T. et al. Overlapping in short motif sequences for binding to human REV7 and MAD2 proteins. *Genes Cells* **15**, 281–296 (2010).

Acknowledgements We thank J. Young (Repare Therapeutics), M. Bashkurov (NBCC, LTRI), L. Kesselman (LTRI), S. Rossi (LTRI) and M. Rother (LUMC); and R. Greenberg (University of Pennsylvania) for the U2OS–FokI cells. We acknowledge that J. Lukas first proposed the name shieldin. S.M.N. is funded by a research fellowship from the Dutch Cancer Society (KWF). S.Ad. is a Banting post-doctoral fellow. D.S. is funded by a post-doctoral research fellowship from CIHR. H.v.A. is funded by the European Research Council (CoG-50364). T.H. is supported by MD Anderson Cancer Center Support Grant P30 CA016672 and the Cancer Prevention Research Institute of Texas (CPRIT/RR160032). Work in the A.M. laboratory is funded through CIHR grant PJT153307. A.-C.G. is funded by CIHR grant FDN143301. Work in the C.J.L. laboratory is funded by Programme Grants from Cancer Research UK (CRUK/A14276) and Breast Cancer Now (CTR-Q4-Y2). Work in the S.R. and J.J. laboratories is funded by the Dutch Cancer Society (KWF 2014-6532), the Netherlands Organization for Scientific Research (VICI 91814643 and a National Roadmap grant for Large-Scale Research Facilities to J.J.), the Swiss National Science Foundation (310030_156869 to S.R.), the European Research Council (CoG-681572 to S.R. and SyG-319661 to J.J.). D.D. and A.C.-G. are Canada Research Chairs (Tier I). D.D. is funded by CIHR grant FDN143343, Canadian Cancer Society (CCS grant #705644) and OICR grant OICR-OC-TRI.

Author contributions S.M.N. initiated the project in the D.D. laboratory, performed the RPE1 *BRCA1*^{KO} screen, validated hits and performed functional follow up. S.Ad. performed validation and localization studies. D.S. carried out structure-function analyses and DNA binding studies. M.B. performed mouse tumour model studies. S.J.P. analysed the SUM149PT talazoparib screen and validated hits. A.K.L. generated knockouts in CH12F3-2 cells, carried out class switch recombination assays and analysed AID levels. M.O. performed the ionizing radiation screen and validation. A.A.-Q. carried out immunopurifications and helped with immunofluorescence quantification. N.M. performed class switch recombination, traffic light reporter and RPA phosphorylation assays. M.Z. carried out the SUM149PT olaparib screen and generated the RPE1 *BRCA1*^{KO} cells. S.An. examined *Shld1* and *Shld2* mutations in mouse embryonic stem cells. D.B.K., I.B., F.S., J.F. and R.B. performed the SUM149PT talazoparib screen and validation. I.B. and F.S. developed SUM149PT knockout cells. F.S. carried out SUM149PT epistasis experiments. A.S. performed the REV7 IP-MS. S.L. initiated work on REV7. R.K.S. generated vectors and edited the manuscript. M.M.M. generated IP-MS cell lines. A.Mc. carried out screen analysis, cloning and TIDE. T.G.d.R. performed shieldin qPCR. Z.-Y.L. carried out IP-MS for SHLD1–3 under supervision of A.-C.G. T.H. analysed screen data and built the TKOv2 library with J.M. A.Ma. supervised A.K.L. J.J. and S.R. supervised M.B. and S.An. D.D. and H.v.A. supervised S.M.N. C.J.L. supervised S.J.P., D.B.K., F.S., I.B., J.F. and R.B. S.M.N. and D.D. wrote the manuscript with input from the other authors.

Competing interests D.D. and T.H. are advisors to Repare Therapeutics. C.J.L. is a named inventor on patents describing the use of PARP inhibitors (DK3044221 (T3), ES2611504 (T3), US2014378525 (A1), WO2008020180 (A2), WO2009027650 (A1)) and stands to gain from their use as part of the ICR 'Rewards to Inventors' scheme.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0340-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0340-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Plasmids. DNA corresponding to sgRNAs was cloned into pX330 (Addgene: 42230, Cambridge), LentiGuide-Puro (Addgene: 52963), LentiCRISPRv2 (Addgene: 52961), or a modified form in which Cas9 was replaced by NLS-tagged GFP or mCherry using *Age*I and BamHI (designated as LentiGuide-NLS–GFP or –mCherry), as described[29,30]. Sequences of the sgRNAs used in this study are included in Supplementary Table 5. Coding sequences of C20orf196 and the short isoform of FAM35A were obtained from the ORFeome collection (http://horfdb.dfci.harvard.edu/), archived in the Lunenfeld-Tanenbaum Research Institute's OpenFreezer[31]. The complete coding sequence of the long isoform of FAM35A was generated by combining a synthesized fragment (GeneArt, Regensburg) corresponding to the long isoform C terminus using an internal KpnI restriction site. The coding sequence of CTC534A2.2 was generated by gene synthesis (GeneArt). The coding sequences were PCR amplified using AscI and ApaI flanking primers and cloned into pcDNA5-FRT/TO-eGFP and pcDNA5-FRT/TO-Flag to obtain N-terminally tagged FAM35A, C20orf196 and CTC534A2.2. pGLUE-HA-Strep-FAM35A was generated by PCR amplification of the long isoform of FAM35A and cloning into pGLUE (Addgene: 15100) using AscI and NotI. To generate FAM35A fragments and mutants, standard protocols for primer-directed mutagenesis or self-ligation of truncated PCR-products were used. To generate pcDNA5-FRT/TO-V5-CTC534A2.2, eGFP was replaced by a V5 tag in the cloning vector pcDNA5-FRT/TO-eGFP using KpnI and AscI restriction enzymes after which the coding sequence for CTC534A2.2 was PCR-amplified and inserted into pcDNA5-FRT/TO-V5-MCS using AscI and XhoI restriction enzymes.

To generate RNF8–FHA fusions, the N terminus of RNF8 (amino acids 1–160) was PCR amplified from pcDNA3-RNF8-FHA(1–160)-RNF168 with flanking AscI sites and inserted into pcDNA5-FRT/TO-eGFP-FAM35A. eGFP (FHA)fusions of FAM35A were introduced into pCW57.1 (Addgene: 41393) by Gateway cloning using the pDONR221 donor vector. FAM35A amino acid substitution mutations and deletions were introduced by site directed mutagenesis and deletion PCR, respectively.

The REV7 coding sequence was obtained from the ORFeome collection and was cloned into the pDEST-FRT/TO-eGFP vector using Gateway cloning and into the pcDNA5-FRT/TO-Flag vector by PCR amplification. The N-terminal 967 residues of RIF1 were amplified by PCR and cloned into the pDONR221 vector using Gateway technology. The fragment was then integrated into the pDEST-mCherry-LacR vector by Gateway cloning. Plasmids for the traffic light reporter system were obtained from Addgene (pCVL-TrafficLightReporter-Ef1a-Puro lentivirus: #31482; pCVL-SFFV-d14GFP-Ef1a-HA-NLS-Sce(opt)-T2A-TagBFP: #32627).

Cell lines and gene editing. 293T and RPE1-hTERT cell lines were obtained from ATCC (Manassas), 293 Flp-In cells were obtained from Invitrogen (Carlsbad) and SUM149PT[32] cells were obtained from Asterand Bioscience (Detroit). U2OS ER-mCherry-LacIFokI-DD cells (U2OS-265, referred to in the text as U2OS–FokI) were a kind gift of R. Greenberg (University of Pennsylvania). All cell lines are routinely authenticated by STR-analysis and tested negative for mycoplasma. 293T, 293 Flp-In, U2OS and RPE1 cells were cultured in high glucose- and GlutaMAX-supplemented DMEM (Gibco, Thermo Fisher Scientific, Waltham) + 1% penicillin–streptomycin (Thermo Fisher Scientific) and 10% heat inactivated fetal calf serum (Wisent, St-Bruno) at 37 °C, 5% CO_2. SUM149PT cells were cultured in Ham's F12 medium (Gibco) supplemented with 5% FCS, 10 mM HEPES, 1 μg/ml hydrocortisone (Sigma-Aldrich, St. Louis), and 5 μg/ml insulin (Sigma-Aldrich) at 37 °C, 5% CO_2. Except for RPE1 clonogenic survival assays, which were performed at 3% O_2, cells were kept under normoxia conditions. Transient transfections of DNA and siRNA were performed using Lipofectamine 2000 (Thermo Fisher Scientific), PEI (Sigma-Aldrich) or calcium phosphate and Lipofectamine RNAiMAX, respectively (Thermo Fisher Scientific). siRNA efficiency was analysed by qPCR and immunoblotting. Stable integration of Flag–C20orf196/FAM35A/REV7 with the Flp-In system was achieved by co-transfection of the pcDNA5-FRT/TO plasmid with the recombinase vector pOG44 (Thermo Fisher Scientific) and hygromycin selection for integration. Lentiviral particles were produced in 293T cells by co-transfection of the targeting vector with vectors expressing VSV-G, RRE and REV using calcium phosphate or PEI (Sigma-Aldrich). Viral transductions were performed in the presence of 4-8 μg/μL polybrene (Sigma-Aldrich) at an MOI <1, unless stated otherwise. Transduced RPE1 cells were selected by culturing in the presence of 15 μg/ml puromycin. For the BRCA1-deficient mouse cell experiments, all experiments were performed using virus produced with the LentiCRISPRv2 backbone (see Supplementary Table 5) and cells were infected using polybrene (8 μg/ml). The medium was refreshed after 12 h and transduced cells were selected with puromycin.

The generation of RPE1-hTERT $TP53^{-/-}BRCA1^{KO}$ Cas9 cells has been described elsewhere[33]. *REV7*, *53BP1*, *FAM35A* and *C20orf196* gene knockouts were generated by electroporation of LentiGuide or LentiCRISPRv2 vectors using a Lonza Amaxa II Nucleofector (Basel) (for sgRNA sequences used see Supplementary Table 5; *REV7* sgRNA1, *FAM35A* sgRNA2 and *C20orf196* sgRNA1 were used for clonal knockout generation). Twenty-four hours after transfection, cells were selected for 24–48 h with 15 μg/ml puromycin, followed by single clone isolation. Triple knockout cell lines of *TP53*, *BRCA1* and *53BP1* were created by mutating *BRCA1* from the $TP53^{-/-}53BP1^{KO}$ double knockout cell line. Triple knockout cell lines of *TP53*, *BRCA1* and *REV7*, *FAM35A* or *C20orf196* were created by mutating *REV7*, *FAM35A* or *C20orf196* in the $TP53^{-/-}BRCA1^{KO}$ cells. Loss of protein(s) was verified by immunoblotting when antibodies were available. Gene mutations were further confirmed by PCR amplification and TIDE analysis[34] (for primers used for genomic PCR, see Supplementary Table 6, for genomic editing information, see Supplementary Table 2).

To generate SUM149PT *53BP1*, *PARP1* or *C20orf196* knockout populations of cells, SUM149PT-doxCAS9 cells were treated with doxycycline for 24 h at 1 μg/ml before transfection with EditR crRNA (Dharmacon, Lafayette). Transfection of guides 53BP1_5_1, 53BP1_5_3, PARP1_5_2, PARP1_5_4, C20orf196_5-1, C20orf196_5-2, C20orf196_5-3 and C20orf196_5-5 (see Supplementary Table 5) was performed at a concentration of 20 nM (crRNA:tracrRNA) in the presence of doxycycline (1 μg/ml) using Lipofectamine RNAiMAX in 48-well plates (35,000 cells per well). The following day cells were split 1:3, fed 24 h later with medium supplemented with 50 nM talazoparib (without doxycycline) and kept in batch culture or further split to generate single cell colonies. Drug-containing medium was replenished every 3–4 days until PARP inhibitor resistant pools or clones were established. Clones were subsequently picked, expanded and validated by genomic PCR and sequence analysis (for primers used, see Supplementary Table 6, for genomic editing information, see Supplementary Table 2). Four SUM149PT $C20orf196^{KO}$ clones with mutations were chosen for further experimentation: clone A (C20orf196 5-1-C1), clone B (C20orf196 5-1-C2), clone C (C20orf196 5-3-C5) and clone D (C20orf196 5-5-C4).

To generate $53BP1^{KO}$ double mutant clones, SUM149PT *C20orf196* clones A and D were infected with a lentivirus expressing an sgRNA targeting *TP53BP1* or a non-targeting control sgRNA (for sequences, see Supplementary Table 5) in medium containing 1 μg/ml doxycycline. Forty-eight hours after infection, puromycin (1 μg/ml) was added to the medium. Selection was maintained for 3 d, until the uninfected control cells were killed. Pools of selection-resistant cells were seeded into 384-well plates for short term survival assays (see below) or subcloned to generate clonal lines.

Mouse embryonic stem cells with a selectable conditional *Brca1* deletion ($Rosa26^{CreERT2/wt}$;$Brca1^{SCo/\Delta}$)[35] were cultured on gelatin-coated plates in 60% buffalo red liver (BRL) cell-conditioned medium supplied with 10% fetal calf serum, 0.1 mM β-mercaptoethanol (Merck, Kenilworth) and 10^3 U/ml ESGRO LIF (Millipore, Burlington) under normal oxygen conditions (21% O_2, 5% CO_2, 37 °C).

The KB1P-G3 2D cell line was previously established from a $Brca1^{-/-}p53^{-/-}$ mouse mammary tumour and cultured as described[14]. In brief, cells were cultured in DMEM/F-12 medium (Life Technologies, Carlsbad) in the presence of 10% FCS, penicillin–streptomycin (Gibco), 5 μg/ml insulin (Sigma-Aldrich), 5 ng/ml epidermal growth factor (Life Technologies) and 5 ng/ml cholera toxin (Gentaur, Kampenhout) under low oxygen conditions (3% O_2, 5% CO_2 at 37 °C).

The KB1P4 3D tumour organoid line was previously established from a $Brca1^{-/-}p53^{-/-}$ mouse mammary tumour and cultured as described[36]. Cells were seeded in Basement Membrane Extract Type 2 (BME, Trevigen, Gaithersburg) on 24-well suspension plates (Greiner Bio-One, Kremsmünster) and cultured in AdDMEM/F12 supplemented with 10 mM HEPES (Sigma-Aldrich), GlutaMAX (Invitrogen), penicillin–streptomycin (Gibco), B27 (Gibco), 125 μM N-acetyl-L-cysteine (Sigma-Aldrich), and 50 ng/ml murine epidermal growth factor (Invitrogen).

CH12F3-2 mutant clones were edited either through transient transfection with pX330 plasmid constructs expressing sgRNAs against *Trp53bp1* (sgRNA: *Trp53bp1*_e6_834, see Supplementary Table 5), *Fam35a*, and *Ctc534a2.2* or by lentiviral lentiCRISPR v2 transduction for *C20orf196*. Double knockout cell lines of *Fam35a* and *Trp53bp1* or *C20orf196* were generated by transient transfection of a pX330 plasmid expressing an sgRNA against *Trp53bp1* or by lentiviral transduction with lentiCRISPRv2 with an sgRNA targeting *C20orf196*.

Antibodies, siRNAs and drugs. An overview of all the antibodies used in this study, including dilution factors, can be found in Supplementary Table 7. The following siRNAs from Dharmacon were used in this study (final siRNA concentration: 10 or 20nM): *53BP1*: siRNA #2 (D-003548-02-0020); *RIF1*: siGENOME *RIF1* siRNA (D-027983-02-0050); *REV7*: siGENOME *MAD2L2* siRNA (M-003272-03-0010); *C20orf196*: SMARTpool: siGENOME *C20orf196* siRNA (M-018767-00-0005); *FAM35A*: SMARTpool: siGENOME *FAM35A* siRNA (M-013761-01-0005); *CTC534A2.2* (custom order): siRNA#1: 5′-GGACAAAACUCAAUCAAAU-3′, siRNA#2: 5′-CAGUAGAUCUAUUGGAGUU-3′, siRNA#3: 5′-CUGGAAGACAUUUGGACAA-3′, siRNA#4: 5′-GCAAGAUAGUUUAAAGGCA-3′ (used as a pool).

The following drugs were used in the course of this study: olaparib (SelleckChem, Houston, or Astra Zeneca, Cambridge), talazoparib (SelleckChem), cisplatin (Sigma-Aldrich), the ATM-inhibitor KU60019 (Sigma-Aldrich), and

© 2018 Springer Nature Limited. All rights reserved.

etoposide (Sigma-Aldrich). Concentration and duration of treatment are indicated in the corresponding figure legends.

Olaparib resistance screens. Viral particles of the TKOv1 sgRNA library were produced as previously described[37]. This library contains 91,320 sgRNA sequences, with a modal number of six sgRNAs per gene. Cas9-expressing cells were infected with an MOI <0.3 and the coverage of sgRNA representation was maintained at >100× (SUM149PT) or >200× (RPE1) (per replicate, if applicable). Twenty-four hours after transduction, transduced cells were selected for 120 h with 10 μg/ml puromycin (RPE1) or 48 h with 3 μg/ml puromycin, followed by 72 h with 0.5 μg/ml puromycin (SUM149PT). Three days after transduction, the transduced cells were split into three technical replicates. Cells were passaged once every three days until nine days after infection, at which time olaparib (16 nM for RPE1 *TP53*$^{-/-}$*BRCA1*^{KO}, 2 μM for SUM149PT) was added to the cells. Olaparib-containing medium was refreshed every 4 d. Cells were collected at 3, 9, 18 and 23 d post-infection (RPE1) or at 3, 9, 19 and 26 d post-infection (SUM149PT) for downstream processing as described[37]. In short, total genomic DNA was isolated from 2×10^7 (t3 sample) or 1×10^7 (later time points) cells using the QIAamp DNA Blood Maxi Kit (Qiagen, Germantown). DNA was precipitated with ethanol and sodium chloride and reconstituted in EB buffer (10 mM Tris-HCl pH 7.5). sgRNA sequences were PCR-amplified using primers harbouring Illumina TruSeq adapters with i5 and i7 barcodes, and the resulting libraries were sequenced on a Illumina NextSeq500 (San Diego) using parameters previously described[37]. Analysis was performed using model-based analysis of genome-wide CRISPR–Cas9 knockout (MAGeCK) version 0.5.3[38], in conjunction with Python v3.5.1 on a Mac OS X El Capitan operating system. Non-treated samples collected at day 9 after transduction were compared to treated samples collected at day 23 (RPE1) or day 26 (SUM149PT). The positive score for each gene was calculated by using the 'run' function with the following arguments: mageck run -l /path/to/TKOv1_library/ -n 08-02-2017_141703–sample-label test,CTRL -t 1 -c 0–fastq /path/to/fastq1 /path/to/fastq2.

Ionizing radiation dropout screen and TKOv2 library. The TKOv2 lentiviral CRISPR library was used for whole-genome CRISPR knockout screening. To design TKOv2, all possible 20-mer sequences upstream of NGG PAM sites were collected where the SpCas9 double-strand break would occur within a coding exon (defined by hg19/Gencode v19 'appris_principal,' 'appris_candidate_longest,' or 'appris_candidate' transcript). Guides with 40–75% GC content were retained and further filtered to exclude homopolymers of length ≥ 4, SNPs (dbSNP138), and relevant restriction sites, including BsmI (GAATCG) and BsmBI (CGTCTC). Candidate gRNA + PAM sequences were mapped to hg19 and sequences with predicted off-target sites in exons or introns, or sequences with more than two predicted off-target sites (with up to two mismatches) in any location, were discarded. Remaining guides were scored using a 'sequence score table' as previously described[40]. Four guides per gene were selected, with a bias towards high sequence scores and maximal coverage across exons (that is, moderate-scoring guides targeting different exons were preferred to high-scoring guides targeting the same exon). The final library contains 70,555 gRNA targeting 17,942 protein-coding genes, as well as 142 sequences targeting LacZ, luciferase, and eGFP. Oligonucleotide sequences were ordered from CustomArray (Bothell), PCR-amplified, and cloned into the pLCKO vector as previously described[37].

RPE1-hTERT *TP53*$^{-/-}$ Cas9-expressing cells were transduced with the lentiviral TKOv2 library (see below) at a low MOI (~0.35) and puromycin-containing medium was added the next day to select for transductants. Selection was continued until 72 h post-transduction, which was considered the initial time point, t0. To identify ionizing radiation sensitizers, the negative-selection screen was performed by subculturing at days 3 and 6 (t3 and t6), at which point the cultures were split into two populations. One was left untreated while the second was treated with 3 Gy of ionizing radiation using a Faxitron X-ray cabinet (Faxitron, Tucson) every 3 d after day 6. Cell pellets were frozen at day 18 for gDNA isolation. Screens were performed in technical duplicates and library coverage of ≥ 375 cells per sgRNA was maintained at every step. gDNA from cell pellets was isolated using the QIAamp Blood Maxi Kit (Qiagen) and genome-integrated sgRNA sequences were amplified by PCR using the KAPA HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington). i5 and i7 multiplexing barcodes were added in a second round of PCR and final gel-purified products were sequenced on Illumina NextSeq500 systems to determine sgRNA representation in each sample. DrugZ[41] was used to identify gene knockouts which were depleted from ionizing radiation-treated t18 populations but not depleted from untreated cells.

Two-colour competitive growth assay. Twenty thousand cells were infected at an MOI of ~1.2 to ensure 100% transduction efficiency with either virus particles of NLS–mCherry LacZ–sgRNA or NLS–GFP GOI–sgRNA. Ninety-six hours after transduction, mCherry- and GFP-expressing cells were mixed 1:1 (2,500 cells + 2,500 cells) and plated with or without olaparib (16 nM) or etoposide (100 nM) in 12-well format. During the course of the experiment, cells were subcultured when near-confluency was reached. Olaparib- or etoposide-containing medium was replaced every three days. Cells were imaged for GFP- and mCherry signal the day of initial plating ($t = 0$) and on days 3, 6, 9, 12, 15 and 18 (olaparib), or, in a separate set of experiments, on days 5, 10, 15 and 20 (etoposide). Cells were imaged using the automatic InCell Analyzer (GE Healthcare Life Sciences, Marlborough) with a 4× objective. Segmentation and counting of the number of GFP-positive and mCherry-positive cells was performed using an Acapella script (PerkinElmer, Waltham). Efficiency of indel formation was analysed by performing PCR amplification of the region surrounding the sgRNA sequence and TIDE analysis on DNA isolated from GFP-expressing cells 9 d post-transduction.

Mass spectrometry. Following 24 h of doxycycline-induction of stably integrated 293 FLP-IN cells (expressing Flag, Flag–FAM35A, Flag–REV7, Flag–C20orf196, Flag–CTC-534A2.2), cell pellets from two 150-mm plates were lysed in 50 mM HEPES-KOH (pH8.0), 100 mM KCl, 2 mM EDTA, 0.1% NP-40, 10% glycerol and affinity-purified using Flag–M2 magnetic beads (Sigma-Aldrich). Subsequently, digestion with trypsin (Worthington, Columbus) was performed on-beads. All immunoprecipitations were performed in biological replicates (three for CTC-534A2.2, five for C20orf196 and six for FAM35A and REV7).

For liquid chromatography–tandem mass spectrometry analysis, peptides were reconstituted in 5% formic acid and loaded onto a 12–15-cm fused silica column with pulled tip packed with C18 reversed-phase material (Reprosil-Pur 120 C18-AQ, 3 μm, Dr. Maisch, Ammerbuch-Entringen). Peptides were analysed using an LTQ-Orbitrap Velos (Thermo Scientific) or a 6600 Triple TOF (AB SCIEX, Framingham) coupled to an Eksigent NanoLC-Ultra HPLC system and a nano-electrospray ion source (Proxeon Biosystems, Thermo Fisher Scientific). Peptides were eluted from the column using a 90–100-min gradient of acetonitrile in 0.1% formic acid. Tandem mass spectrometry spectra were acquired in a data-dependent mode for the top 10 most abundant ions using collision-induced dissociation. After each run, the column was washed extensively to prevent carry-over.

Mass spectrometry data extraction and interaction scoring was performed essentially as described previously[42]. In short, raw mass spectrometry files were converted to mzXML and analysed using the iProphet pipeline[43], implemented within ProHits[44]. The data were searched against the human and adenovirus complements of the Uniprot (forward and reverse) database (Version 2017_09; reviewed Swiss-Prot entries only), to which common epitope tags were added as well as common contaminants (common contaminants are from the Max Planck Institute, http://141.61.102.106:8080/share.cgi?ssid=0f2gfuB, and the Global Proteome Machine, http://www.thegpm.org/crap/index.html; 85,393 entries were searched). Mascot and Comet search engines were used with trypsin specificity (2 missed cleavages allowed) and deamidation (NQ) and methylation (M) as variable modifications. Charges +2, +3 and +4 were allowed with a parental mass tolerance of maximum 12 ppm and a fragment bin tolerance of 0.6 Da selected for Orbitrap instruments, while 35 ppm and 0.15 Da were allowed for the TripleTOF 6600. For subsequent SAINT analysis (see below), only proteins with an iProphet protein probability ≥ 0.95 were considered, corresponding to an estimated protein FDR of ~0.5%.

Interactions were analysed with SAINTexpress (v3.6.1)[45,46]. SAINT probability scores were computed independently for each replicate against eight biological replicate analyses of the negative control (Flag alone; controls were 'compressed' to six virtual controls to increase robustness as described[47]) and the average probability (AvgP) of the best three out of three (CTC534A2.2), five out of five (C20orf196) or six (FAM35A, REV7) biological replicates for each bait was reported as the final SAINT score. Preys with an estimated FDR ≤ 1% were considered true interactors (AvgP ≥ 0.91). The entire dataset, including the peptide identification and complete SAINTexpress output was deposited as a complete submission in ProteomeXchange through the partner MassIVE housed at the Center for Computational Mass Spectrometry at University of California, San Diego (UCSD; http://massive.ucsd.edu). Data are available at MassIVE (ftp://massive.ucsd.edu/MSV000082207). Unique accession numbers are MSV000082207 and PXD009313, respectively. Data can also been viewed at the prohits website (prohits-web.lunenfeld.ca) under dataset 29: Durocher laboratory. Data in Fig. 1f is represented using Cytoscape, using analyses with an FDR ≤ 1 or 5%.

Immunoprecipitation. 293T cells (1×10^7) were transfected with pcDNA5.1-FRT/TO-Flag-c20orf196 (10 μg), -GFP-REV7 (2 μg), -V5-CTC534A2.2 (14 μg) and pGLUE-HA-Strep-FAM35A (14 μg) or empty vectors using a standard calcium phosphate or PEI protocol. After 48 h, cells were washed with PBS, scraped, and lysed in 1 ml of lysis buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 2 mM EDTA, 0.5% NP-40, 10mM NaF, 10 mM $MgCl_2$ and 10 U/ml Benzonase (Sigma-Aldrich)) on ice for 30 min. Lysates were centrifuged at 15,000*g* for 5 min at 4 °C, and supernatants were incubated with 100 μl of Streptavidin Sepharose High Performance beads (GE Healthcare) or Dynabeads M-280 Streptavidin magnetic beads (Invitrogen) for 1 h at 4 °C. Beads were washed 5 times with lysis buffer and eluted with 10 mM D-biotin (Invitrogen) in lysis buffer for 2 h at 4 °C. When applicable, the eluate was incubated with 20 μl of GFP-Trap_M resin (Chromotek, Planegg-Martinsried) for 1 h at 4 °C, washed 5 times with lysis buffer and eluted by boiling

© 2018 Springer Nature Limited. All rights reserved.

in sample buffer. Pull-downs and whole cell extracts were loaded onto SDS–PAGE gels, followed by immunoblotting and probing with indicated antibodies. For GFP–CTC534A2.2 immunoprecipitations, an identical GFP-Trap pulldown procedure as above was used. For V5–CTC534A2.2 immunoprecipitations, lysates from one confluent 10-cm dish of 293T cells transfected with 10 μg pcDNA5.1-FRT/TO-V5-CTC534A2.2 vector was incubated with 10 μg/ml anti-V5 antibody (Invitrogen) for 2 h at 4 °C. Subsequently, 50 μl of protein G Dynabeads (Invitrogen) was added to the lysates and incubated for an additional 1 h at 4 °C. Beads were washed 4 times with lysis buffer and boiled in 50 μl 2 × SDS buffer.

Clonogenic survival assays. RPE1-hTERT *TP53*$^{-/-}$ cells were seeded in 10-cm dishes (wild type, 250 cells; *BRCA1*^{KO}*53BP1*^{KO}, 500 cells; *BRCA1*^{KO} or *BRCA1*^{KO}*C20orf196*^{KO}, 1,500 cells; *BRCA1*^{KO}*FAM35A*^{KO}, 750 cells) in the presence of 800 nM cisplatin or 16 nM olaparib or left untreated. Cisplatin dosing lasted 24 h, after which cells were grown in drug-free medium. Olaparib-containing medium was refreshed after 7 d. After 14 d, colonies were stained with crystal violet solution (0.4% (w/v) crystal violet, 20% methanol) and manually counted. Relative survival was calculated for the drug treatments by setting the number of colonies in non-treated controls at 100%.

For *Rosa26*$^{CreERT2/wt}$*Brca1*$^{SCo/\Delta}$ cells, Cre-mediated inactivation of the endogenous mouse *Brca1SCo* allele was achieved by overnight incubation of cells with 0.5 μM 4-hydroxytamoxifen (Sigma-Aldrich). Four days after switching, cells were seeded in triplicate at 10,000 cells per well in 6-well plates for clonogenic survival assays. For experiments with *Rosa26*$^{CreERT2/wt}$*Brca1*$^{SCo/\Delta}$ p53-null cells, cells were plated in the presence of 15 nM olaparib. Cells were stained with 0.1% crystal violet one week later. Clonogenic survival assays with PARPi (olaparib) and ATMi (KU60019) combination treatment were performed as described previously with minor adjustments[42]. Five thousand KB1P-G3 cells were seeded per well into 6-well plates on day 0, and then PARPi, ATMi, both or neither were added. The medium was refreshed every 3 d. On day 6, the ATMi alone and untreated groups were stopped and stained with 0.1% crystal violet, the other groups were stopped and stained on day 9. Plates were scanned with a GelCount (Oxford Optronics, Abingdon). Quantifications were performed by solubilizing the retained crystal violet using 10% acetic acid and measuring the absorbance at 562 nm using a Tecan plate reader (Tecan, Männedorf).

Short-term survival assays. Ten thousand RPE1-hTERT Cas9 *TP53*$^{-/-}$ parental cells and additional mutants (*BRCA1*^{KO} and/or *FAM35A*^{KO}) with or without stable integration of indicated eGFP fusions by viral transduction were seeded in 12-well format with or without 200 nM olaparib (and 1 μg/ml doxycycline if applicable). Medium with olaparib (and doxycycline) was replaced after 4 d, and cells were trypsinized and counted after seven days using an automated Z2 Coulter Counter analyser (Beckman Coulter, Indianapolis).

SUM149PT cells were plated at 500 cells per well in 384-well plates and varying amounts of talazoparib in DMSO were added the following day using an Echo 550 liquid handler (Labcyte, San Jose). After 5 d of growth, cell survival was assayed using CellTiter-Glo according to the manufacturer's protocol (Promega, Madison).

Immunofluorescence. For 53BP1 immunofluorescence, cells were cultured on coverslips and treated with 5 or 10 Gy X-ray irradiation and fixed with 2–4% paraformaldehyde (PFA) 1 h after irradiation. Cells were permeabilized with 0.3% Triton X-100, followed by blocking in 10% goat serum, 0.5% saponin, 0.5% NP-40 in PBS (blocking buffer A). Cells were co-stained using 53BP1 and γH2AX primary antibodies (see Supplementary Table 7) in blocking buffer A for 1.5 h at room temperature, followed by 4 washes in PBS, incubation with appropriate secondary antibodies in blocking buffer A plus 0.8 μg/mL DAPI for 1.5 h at room temperature, and finally four washes in PBS.

For RAD51 immunofluorescence, cells with or without stable integration of eGFP-tagged proteins or sgRNAs via viral transduction were grown on glass coverslips and treated with 10 Gy X-ray irradiation and recovered for 3 to 6 h (as indicated). Cells were fixated and extracted using 1% PFA, 0.5% Triton X-100 in PBS for 20 min at room temperature, followed by a second extraction/fixation using 1% PFA, 0.3% Triton X-100, 0.5% methanol in PBS for 20 min at room temperature. Blocking, primary and secondary antibody incubations (1.5 h at room temperature followed by 4 PBS washes) were performed in BTG buffer (10 mg/mL BSA, 0.5% Triton X-100, 3% goat serum, 1 mM EDTA in PBS) or PBS^{+} (0.5% BSA, 0.15% glycine in PBS).

For REV7 and RIF1 immunofluorescence, cells were grown on glass coverslips, treated with 5 or 10 Gy X-ray irradiation and fixed with 2–4% PFA 1-2 h after irradiation. Cells were permeabilized with 0.3% Triton X-100. For REV7 blocking, primary and secondary antibody incubations (1.5 h at room temperature followed by 4 washes in PBS) were performed in blocking buffer A. For RIF1 blocking, primary and secondary antibody incubations (1.5 h at room temperature followed by 4 washes) were performed in PBG buffer (0.2% cold water fish gelatin (Sigma Aldrich), 0.5% BSA in PBS).

For GFP–shieldin focus and laser stripe analysis, U2OS or RPE1 cells were grown on glass coverslips and either transiently transfected with 1 μg vector expressing GFP–FAM35A or –CTC534A2.2, or virally transduced with GFP–FAM35A-expressing vector. Forty-eight hours after transfection, or 24 h after 0.5 μg/ml doxycyclin induction, cells were treated with 5 Gy X-ray irradiation or micro-irradiated, pre-extracted 10 min on ice with NuEx buffer (20 mM HEPES, pH 7.4, 20 mM NaCl, 5 mM MgCl2, 0.5% NP-40, 1 mM DTT and protease inhibitors) followed by 10 min 2% PFA fixation 1 h post- ionizing radiation or micro-irradiation. Antibody staining and blocking were performed as described above except in PBS + 0.1% Tween-20 + 5% BSA using GFP and γH2AX antibodies.

DAPI (0.8 μg/ml) was included in all experiments to stain nuclear DNA. Coverslips were mounted using Prolong Gold mounting reagent (Invitrogen) or Aqua PolyMount (Polyscience, Warrington). Images were acquired using a Zeiss LSM780 laser-scanning microscope (Oberkochen), a Leica SP8 confocal microscope (Wetzlar) or a Zeiss AxioImager D2 widefield fluorescence microscope. Foci were manually counted.

RAD51 immunofluorescence in KB1P-G3 cells was performed as previously described, with minor modifications[10]. Cells were grown on 8-well chamber slides (Millipore). Ionizing-radiation induced foci were induced by gamma-irradiation (10 Gy) 4 h before sample preparation. Cells were then washed in PBS++ (2% BSA, 0.15% glycine, 0.1% Triton X-100) and fixed with 2% PFA/PBS++ for 20 min on ice. Fixed cells were washed with PBS++ and were permeabilized for 20 min in 0.2% Triton X-100/PBS++. All subsequent steps were performed in PBS++. Cells were washed thrice and blocked for 30 min at room temperature, incubated with the primary antibody for 2 h at room temperature, washed thrice and incubated with the secondary antibody for 1 h at room temperature. Lastly, cells were mounted and counterstained using Vectashield mounting medium with DAPI (H1500, Vector Laboratories, Burlingame).

Scale bars indicated in the figure panels represent 10 μm, unless stated otherwise.

LacR–RIF1 N-terminus and FokI-induced focus formation. For monitoring recruitment of GFP-tagged shieldin subunits to mCherry–LacR–Rif1(1–967) foci, 150,000 U2OS–FokI cells (known also as U2OS-DSB)[48] were seeded on 6-well plates containing glass coverslips without any induction of FokI. Twenty-four hours after seeding, cells were transfected using 1μg of pDEST-mCherry-LacR or pDEST-mCherry-LacR-Rif1(1-967), if applicable, and 0.5–1 μg of GFP fusion expression vectors. Cells were fixed with 4% PFA 24–48 h after transfection. For monitoring the localization of the FAM35A N terminus to Rif1(1–967) foci with siRNA knockdown of other shieldin subunits, an essentially identical protocol was used with the following adjustments: 350,000 U2OS–FokI cells were reverse-transfected with Lipofectamine RNAiMAX-siRNA (10 nM) complex. Twenty-four hours after siRNA transfection, the mCherry–LacR and GFP fusion plasmids were transfected. Cells were fixed with 4% PFA 48 h after DNA transfection. For monitoring recruitment of GFP-tagged shieldin subunits to DSBs at the LacO array, FokI stabilization and nuclear translocation was induced by treating cells with 0.1 μM Shield1 (Clontech, Mountain View) and 10 μg/ml hydroxytamoxifen for 4 h.

ImageJ (https://imagej.nih.gov/ij/) was used to quantify foci in the U2OS–FokI system. An mCherry focus and DAPI nuclear signal were used to generate masks. The average GFP fluorescence or immunofluorescence intensity in the mCherry focus mask was divided by the corresponding average nuclear intensity, and the ratio is reported. Cells displaying a ratio of focus/nuclear average intensity >3 are defined as containing a focus.

Microirradiation. For laser microirradiation, virally transduced RPE1 cells expressing the indicated eGFP-tagged proteins were grown on glass coverslips and transfected with siRNAs. Forty-eight hours after transfection, protein expression was induced using 0.5 μg/ml doxycycline, and 24 h later cells were presensitized with 1 μg/ml Hoechst for 15 min at 37 °C. DNA damage was introduced with a 355-nm laser (Coherent, Santa Clara, 40mW) focused through a Plan-Apochromat 40× oil objective to yield a spot size of 0.5–1 mm using a LSM780 confocal microscope (Zeiss) and the following laser settings: 100% power, 1 iteration, frame size 128 × 128, line step 7, pixel dwell: 25.21 μs.

Traffic light reporter assay. Cells were infected with pCVL.TrafficLightReporter.Ef1a.Puro lentivirus at a low MOI (0.3–0.5) and selected with puromycin (15 μg/μl). Cells (7×10^5) were nucleofected with 5 μg of pCVL.SFFV.d14GFP.Ef1a.HA.NLS.Sce(opt).T2A.TagBFP plasmid DNA in 100 μl of electroporation buffer (25 mM Na_2HPO_4 pH 7.75, 2.5 mM KCl, 11 mM $MgCl_2$), using program T23 on a Nucleofector 2b (Lonza). After 72 h, GFP and mCherry fluorescence was assessed in BFP-positive cells using a Fortessa X-20 (BD Biosciences, San Jose) flow cytometer.

Phospho-RPA immunoblotting. For phospho-RPA staining, CH12 cells were left untreated, or were treated with 25 Gy of ionizing radiation using a Faxitron X-ray cabinet, and were then collected by centrifugation 3 h later. Pellets were lysed on ice for 10 min in high salt lysis buffer (50 mM Tris-HCl pH 7.6, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100, 1 mM DTT, 1× EDTA-free protease inhibitor cocktail (Roche, Basel)), cleared by centrifugation at 20,000*g* for 10 min at 4 °C, and quantified by bicinchoninic acid assay (BCA; Pierce, Thermo Fisher Scientific). Equal amounts of whole-cell extracts were separated by SDS–PAGE on 4–12% Bis-Tris

© 2018 Springer Nature Limited. All rights reserved.

gradient gels (Invitrogen), transferred to nitrocellulose and immunoblotted for pRPA32 (S4/S8).

Mouse mammary tumour models. All animal experiments were approved by the Animal Ethics Committee of The Netherlands Cancer Institute (Amsterdam) and performed in accordance with the Dutch Act on Animal Experimentation (November 2014). KB1P4 tumour organoids were transduced using spinoculation as previously described[49]. NMRI-nude female mice were purchased from Janvier Laboratories and used for transplantation studies at the age of 6–9 weeks. A power analysis was performed to calculate that a minimum of 8 mice per group were needed to achieve a power of 0.8 (two-sided test, alpha = 0.05).Tumour organoids were allografted in mice as previously described with minor adjustments[36]. In brief, tumour organoids were collected, incubated with TripLE at 37 °C for 5 min, dissociated into single cells, washed and embedded in a 1:1 mixture of tumour organoid culture medium and Basement Membrane Extract (Trevigen) in a cell concentration of 10^4 cells per 40 μl. Subsequently, 10^4 cells were injected in the fourth right mammary fat pad of NMRI nude mice. Mammary tumour size was determined by caliper measurements and tumour volume was calculated ($0.5 \times \text{length} \times \text{width}^2$). Treatment of tumour-bearing mice was initiated when tumours reached a size of 50–100 mm^3. Mice were randomly allocated into the untreated ($n = 8$) or olaparib treatment group ($n = 8$). Olaparib was administered in a blinded fashion at 100 mg/kg intraperitoneally for 80 consecutive days. Animals were killed with CO_2 when the tumour reached a volume of 1,500 mm^3. The tumour was collected, fixed in formalin for histology and several tumour pieces were collected for DNA analysis.

Class switch recombination assays. To induce switching in CH12F3-2 murine B cell lymphoma cells, 2×10^5 cells were cultured in CH12 medium supplemented with a mixture of IL4 (10 ng/mL, R&D Systems #404-ML-050, Minneapolis), TGFβ (1 ng/ml, R&D Systems #7666-MB-005) and anti-CD40 antibody (1 μg/ml, #16-0401-86, eBioscience, Thermo Fisher) for 48 h. Cells were then stained with anti-IgA-PE and fluorescence signal was acquired on an LSR II or Fortessa X-20 flow cytometer (BD Biosciences). To probe AID levels in the stimulated cells, immunoblotting was performed on total cell lysates using anti-AID and anti-β-actin antibodies (Supplementary Table 7). Band quantification was analysed by ImageJ.

Plasmid integration assay. Two hundred thousand RPE1 cells were seeded into 6-well plates and 24 h later transfected with 2 μg of BamHI/EcoRI-linearized peGFP-c1 using PEI. Seventy-two hours after transfection, cells were seeded for colony formation into 10-cm dishes in the presence (50,000 cells per dish) or absence (500 cells per dish) of 600 μg/ml G418. At this point, transfection efficiency was analysed by measuring GFP-positivity using flow cytometry. Medium with G418 was refreshed every 3 d. Fourteen days after seeding, colonies were stained with crystal violet solution and manually counted. NHEJ efficiency was calculated according to the following formula:

$$\frac{\text{Percentage of surviving colonies on selection}}{\text{Percentage of surviving colonies without selection}} \times \text{percentage of transfected cells}$$

The data shown for the different knockout clones in Extended Data Fig. 8e represent NHEJ efficiency as calculated with the above formula, followed by normalization to wild-type cells (for which NHEJ efficiency is set to 100%).

DNA binding assays. Shieldin proteins were isolated using the immunoprecipitation protocol described above with the following modifications. 293T cells were transfected with pGLUE- FAM35A(421-904), the indicated mutants of this construct, or the empty pGLUE Strep/HA-tagging vector and pcDNA5.1-FRT/TO-Flag-C20orf196 in a 2:1 ratio for a total of 10 μg per 10 cm dish. Complexes were immunoprecipitated as described, except using a reduced NP-40 detergent concentration (0.1%) for the last two washes and elution buffer. Eluted proteins were concentrated using Amicon Ultra 0.5 ml 10K centrifugal filter units (Millipore). Concentrations of isolated proteins were estimated by SDS–PAGE and Coomassie staining, followed by comparison to a standard curve of known bovine serum albumin (BSA) concentrations measured by fluorescence in the 700-nM channel of the Odyssey imager (LI-COR). A radiolabelled ssDNA probe was prepared by T4 polynucleotide kinase (New England Biolabs, Ipswich) phosphorylation of HPLC-purified 59-nt DNA oligonucleotide (BioBasic, Markham); TACGTTAGTATGCGTTCTTCCTTCCAGAG GTTTTTTTTTTTTTTTTTTTTTTTTTTTTT) using [γ-^{32}P]ATP (3,000 Ci/mmol, 10 mCi/ml; Perkin-Elmer, Woodbridge). Unlabelled competitors were prepared using the same oligonucleotide sequence alone or hybridized to the complementary sequence (AAAAAAAAAAAAAAAAAAAAAAAAAAAAAACCTCTGGAAGGAAGAACGCATACTAACGTA) by heating at 80 °C for 10 min and gradual cooling to room temperature overnight.

For electrophoretic mobility shift assays, 20 nM of labelled ssDNA probe was incubated with purified proteins for 20 min in the elution buffer with the addition of 1 mM DTT and 1 mg/ml BSA at room temperature. Glycerol was then added to a final concentration of 8.3% and resolved on 6% acrylamide-TAE gels. Gels were adhered onto blotting paper (VWR) and enclosed in plastic wrap. Gels were exposed to a storage phosphor screen (GE Healthcare) and visualized using a Typhoon FLA 9500 biomolecular imager (GE Healthcare). Dissociation constant (K_d) was determined in GraphPad Prism from nonlinear regression analysis assuming single-site specific binding of saturation titration experiments, defining all signal above the free probe band to be bound probe, as measured in ImageQuant TL(GE Healthcare). The fraction of probe bound is defined as:

$$\frac{\text{Signal of bound probe}}{\text{Signal of bound probe} + \text{signal of free probe}}$$

and the concentration of unbound FAM35AC–C20orf196 (referred in the text as SHLD2-C–SHLD1) complex is calculated by multiplying the fraction of probe bound by the initial concentration of ssDNA probe, and subtracting this from the initial concentration of SHLD2-C–SHLD1, given the assumption of 1:1 binding.

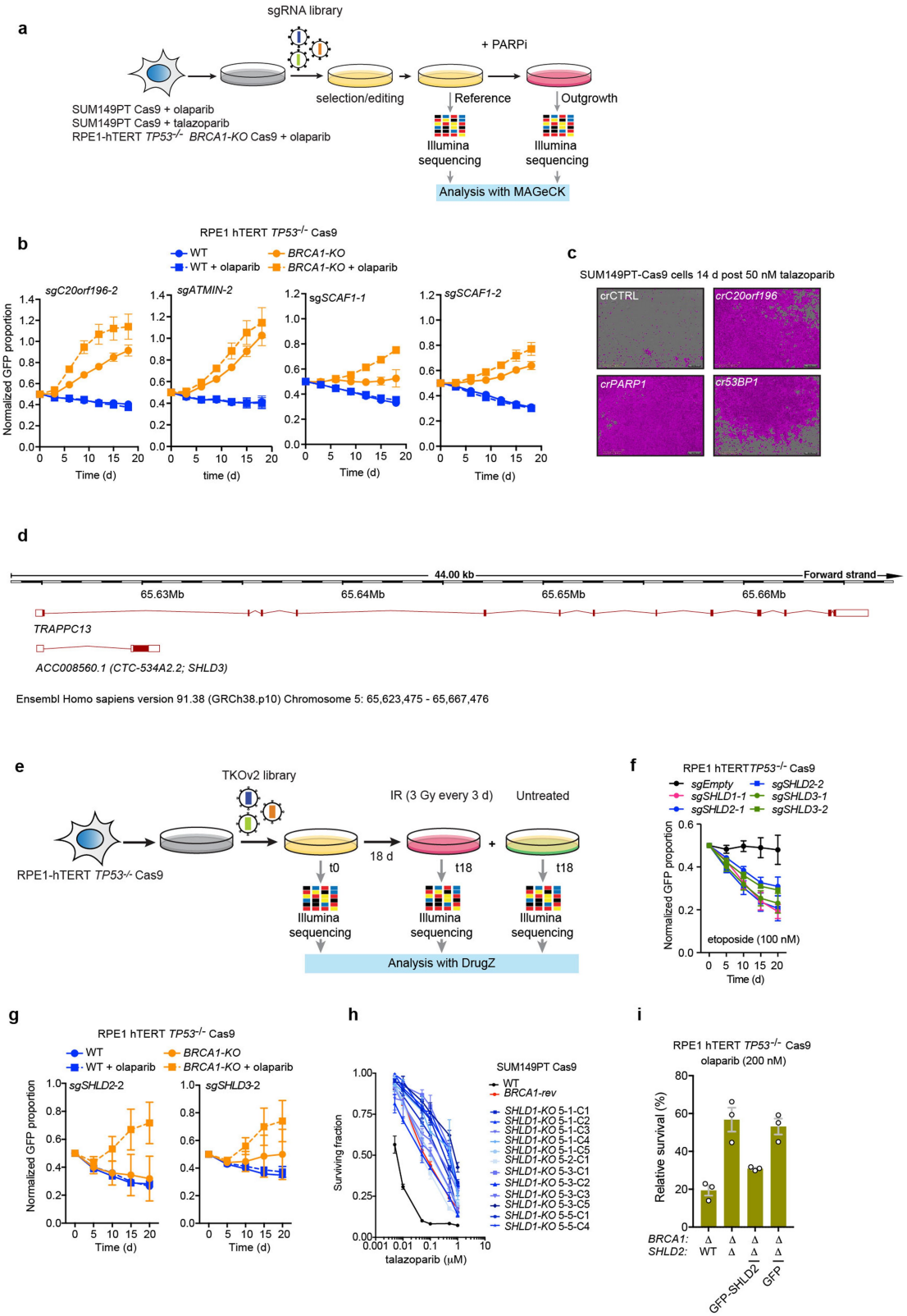
Statistical analysis. All data are represented as individual replicates and replicate number, mean and error bars are explained in the figure legends. The statistical tests we used (all of which were common tests) and resulting *P* values are indicated in the figure legends and/or figure panels and have been generated using GraphPad Prism software.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All source data represented in the graphs displayed in this article are available online (Supplementary Data 1–12). Uncropped western blots can be found online as Supplementary Fig. 1. Data of the CRISPR–Cas9 screens are included as Supplementary Table 1 (PARPi positive selection screens) or Supplementary Table 4 (ionizing radiation sensitivity dropout screen). IP-MS data (Supplementary Table 3) are available at MassIVE (ftp://massive.ucsd.edu/MSV000082207, with unique accession numbers MSV000082207 and PXD009313). IP-MS data can also been viewed at the prohits website (http://prohits-web.lunenfeld.ca) under dataset 29: Durocher laboratory.

29. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
30. Shalem, O. et al. Genome-scale CRISPR–Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
31. Olhovsky, M. et al. OpenFreezer: a reagent information management software system. *Nat. Methods* **8**, 612–613 (2011).
32. Elstrodt, F. et al. BRCA1 mutation analysis of 41 human breast cancer cell lines reveals three new deleterious mutants. *Cancer Res.* **66**, 41–45 (2006).
33. Zimmermann, M. et al. CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions. *Nature* https://doi.org/10.1038/s41586-018-0291-z (2018).
34. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168 (2014).
35. Bouwman, P. et al. A high-throughput functional complementation assay for classification of BRCA1 missense variants. *Cancer Discov.* **3**, 1142–1155 (2013).
36. Duarte, A. A. et al. BRCA-deficient mouse mammary tumor organoids to study cancer-drug resistance. *Nat. Methods* **15**, 134–140 (2018).
37. Hart, T. et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).
38. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
39. Tzelepis, K. et al. A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. *Cell Reports* **17**, 1193–1205 (2016).
40. Hart, T. et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).
41. Wang, G. et al. Identifying drug–gene interactions from CRISPR knockout screens with drugZ. Preprint at https://doi.org/10.1101/232736 (2017).
42. Tkáč, J. et al. HELB is a feedback inhibitor of DNA end resection. *Mol. Cell* **61**, 405–418 (2016).
43. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **10**, M111.007690 (2011).
44. Liu, G. et al. ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nat. Biotechnol.* **28**, 1015–1017 (2010).
45. Choi, H. et al. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods* **8**, 70–73 (2011).
46. Teo, G. et al. SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J. Proteomics* **100**, 37–43 (2014).
47. Mellacheruvu, D. et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10**, 730–736 (2013).
48. Tang, J. et al. Acetylation limits 53BP1 association with damaged chromatin to promote homologous recombination. *Nat. Struct. Mol. Biol.* **20**, 317–325 (2013).
49. Koo, B. K. et al. Controlled gene expression in primary Lgr5 organoid cultures. *Nat. Methods* **9**, 81–83 (2012).
50. Dréan, A. et al. Modeling therapy resistance in *BRCA1/2*-mutant cancers. *Mol. Cancer Ther.* **16**, 2022–2034 (2017).

© 2018 Springer Nature Limited. All rights reserved.

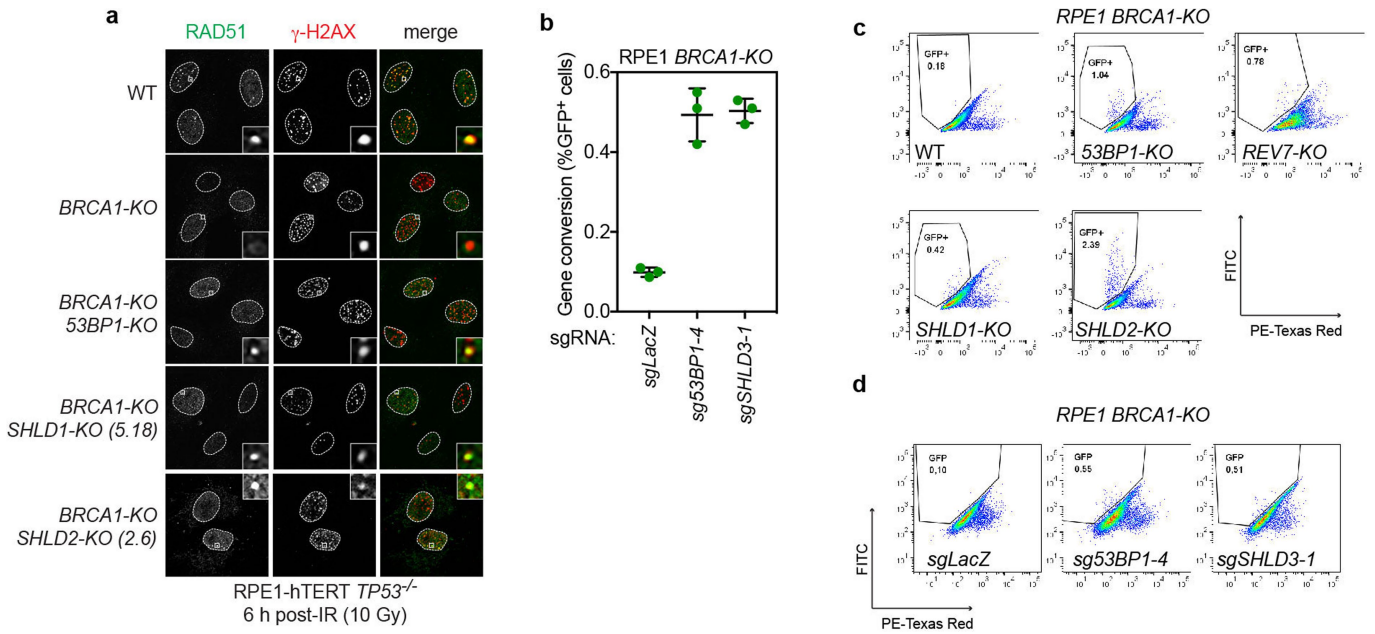


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

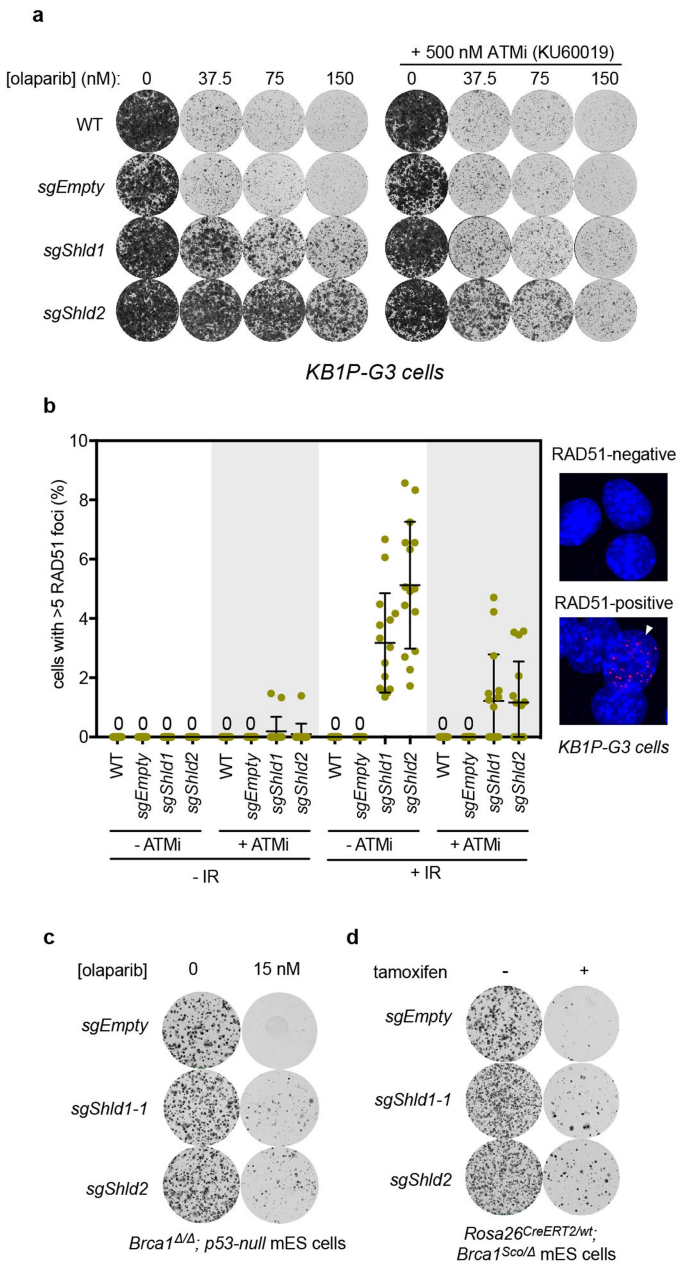
Extended Data Fig. 1 | The identification of the shieldin complex and its role in the response to genotoxic treatments. a, Schematic of the PARPi resistance screens. **b**, Competitive growth assays determining the capacity of the indicated sgRNAs to cause resistance to PARP inhibitors in RPE1 *BRCA1*^{KO} cells. Data are presented as the mean fraction of GFP-positive cells ± s.e.m., normalized to day 0 ($n = 3$, independent viral transductions). Gene-editing efficiencies of the sgRNAs can be found in Supplementary Table 2. Note that we have not been able to obtain TIDE data for the *ATMIN*-targeting sgRNAs. **c**, Representative images of SUM149PT–Cas9 cells transfected with indicated crRNAs (see Methods) and exposed to 50 nM talazoparib for 14 d. Purple colour indicates cells detected by Incucyte live-cell imaging. Scale bar, 100 μm. The data are a representative set of images from two biologically independent experiments. **d**, Screenshot of the genomic locus surrounding human *CTC-534A2.2* taken from ENSEMBL. **e**, Schematic of the screen performed in RPE1-hTERT *TP53*$^{-/-}$ cells stably expressing Cas9 to study genes mediating ionizing radiation-sensitivity. **f**, **g**, Competitive growth assays measuring the capacity of the indicated sgRNAs to cause resistance to etoposide (100 nM) in RPE1 wild-type cells (**f**) or PARPi (16 nM) in RPE1 *BRCA1*^{KO} cells (**g**). Data are presented as the mean fraction of GFP-positive cells ± s.d., normalized to day 0 ($n = 3$, independent viral transductions). Gene-editing efficiencies of the sgRNAs can be found in Supplementary Table 2. **h**, Talazoparib sensitivity in 11 *SHLD1*^{KO} SUM149PT clones obtained after co-transfection of tracrRNA and one of four distinct *SHLD1* crRNAs (5-1, 5-2, 5-3 or 5-5). Each clone was exposed to talazoparib in a 384-well plate format for 5 days. As a comparison, talazoparib sensitivity in parental SUM149PT cells with wild-type SHLD1 (WT) is shown, as is talazoparib resistance in a *BRCA1* revertant subclone (*BRCA1-rev*) of SUM149PT[50]. Bars represent the mean ± s.d. ($n = 4$ biologically independent experiments). ANOVA was performed for each *SHLD1*^{KO} clone versus wild type using Dunnett correction for multiple comparisons, $P < 10^{-15}$. Gene-editing efficiencies can be found in Supplementary Table 2. **i**, *BRCA1*^{KO} and *BRCA1*^{KO}*SHLD2*^{KO} cells were virally transduced with expression vectors for GFP alone or GFP–SHLD2. Sensitivity to olaparib (200 nM) was determined by a short-term survival assay in the presence of 1 μg ml^{-1} doxycycline to induce protein expression. Data are represented as dots for every individual experiment with the bar representing the mean ± s.d. ($n = 3$).

© 2018 Springer Nature Limited. All rights reserved.



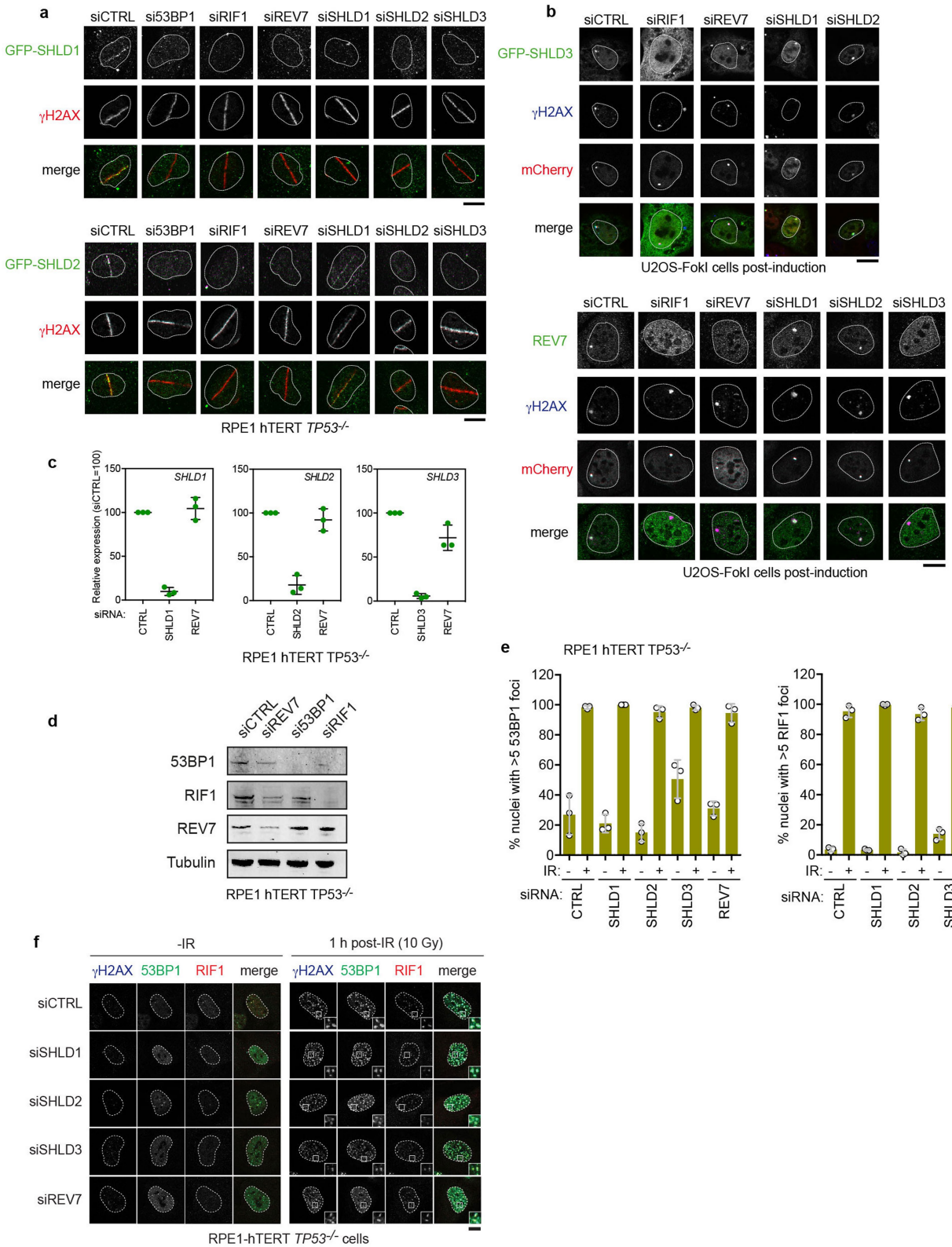
Extended Data Fig. 2 | Shieldin inhibits homologous recombination. **a**, Representative micrographs of RAD51 focus formation in the indicated RPE1 cell lines (data quantified in Fig. 2d, $n \geq 3$). **b**, Traffic light reporter assay testing RPE1 *BRCA1*^{KO} cells virally transduced with sgRNAs targeting *53BP1* or *SHLD3*. Data are represented as dots for individual experiments with the bar representing the mean ± s.d. ($n = 3$). Gene-editing efficiencies of the sgRNAs can be found in Supplementary Table 2. **c**, Representative flow cytometry plots of cells analysed with the traffic light reporter assay (data quantified in Fig. 2e, $n \geq 3$). **d**, Representative flow cytometry plots of cells analysed with the traffic light reporter assay (data quantified in **b**).

© 2018 Springer Nature Limited. All rights reserved.



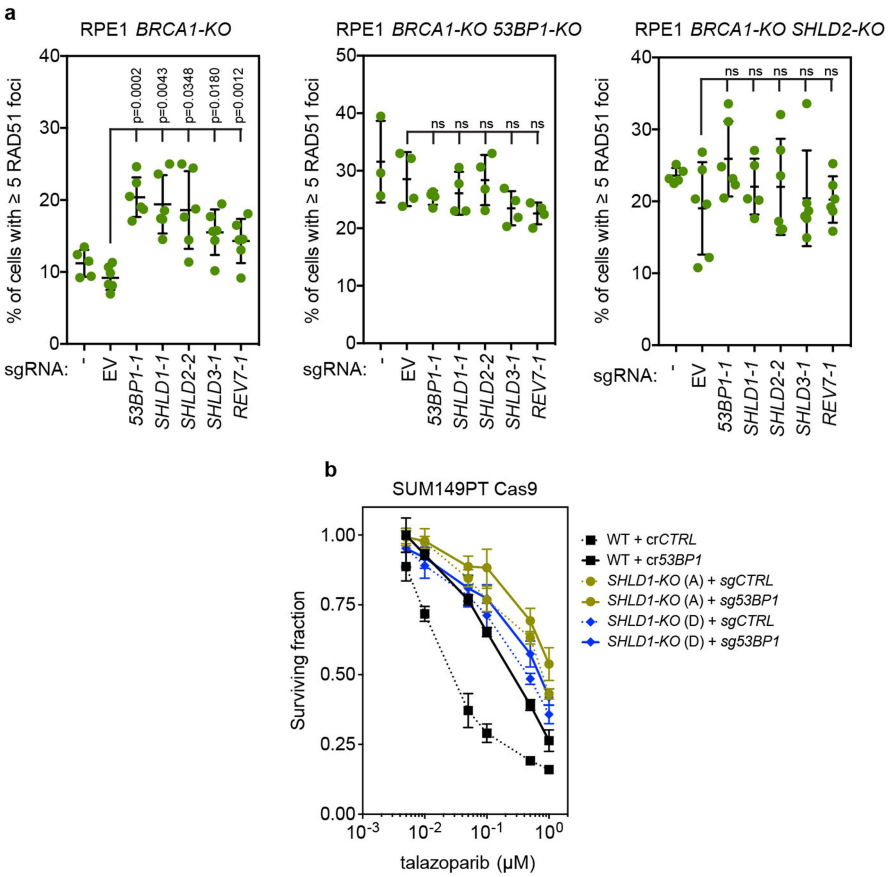
Extended Data Fig. 3 | Mouse shieldin promotes resistance to PARP inhibition in *Brca1*-mutated cells and tumours. a, Clonogenic survival assays of transduced KB1P-G3 cells treated with indicated olaparib doses ± ATM inhibitor (ATMi) KU60019 (500 nM). On day 6, the ATMi alone and untreated groups were stopped and stained with 0.1% crystal violet; the other groups were stopped and stained on day 9. Data shown are representative of 3 biologically independent experiments (with 3 technical replicates each). **b**, Left, quantification of RAD51 focus formation in parental KB1P-G3 ($Brca1^{-/-}$;$Trp53^{-/-}$) cells or KB1P-G3 cells that were transduced with the indicated lentiviral sgRNA vectors. Cells were fixed without treatment or 4 h after irradiation (10-Gy dose). Each data point represents a microscopy field containing a minimum of 50 cells; the bar represents the mean ± s.d. ($n = 15$). Right, representative micrographs of RAD51-negative and RAD51-positive cells (the latter is indicated by an arrowhead). DNA was stained with DAPI. **c**, Clonogenic survival assay of $Rosa26^{CreERT2/wt}$;$Brca1^{\Delta/\Delta}$;*p53*-null mouse embryonic stem cells virally transduced with the indicated sgRNA and treated without or with 15 nM olaparib for 7 d. Gene-editing efficiencies of the sgRNAs can be found in Supplementary Table 2. Data shown are representative of 3 biologically independent experiments (with ≥ 2 technical replicates each). **d**, Clonogenic survival assay of $Rosa26^{CreERT2/wt}$;$Brca1^{Sco/\Delta}$ mouse embryonic stem cells virally transduced with the indicated sgRNA and treated without or with 0.5 μM tamoxifen to induce BRCA1 depletion. Gene-editing efficiencies of the sgRNAs can be found in Supplementary Table 2. Data shown are representative of 2 biologically independent experiments (with 3 technical replicates each).

© 2018 Springer Nature Limited. All rights reserved.



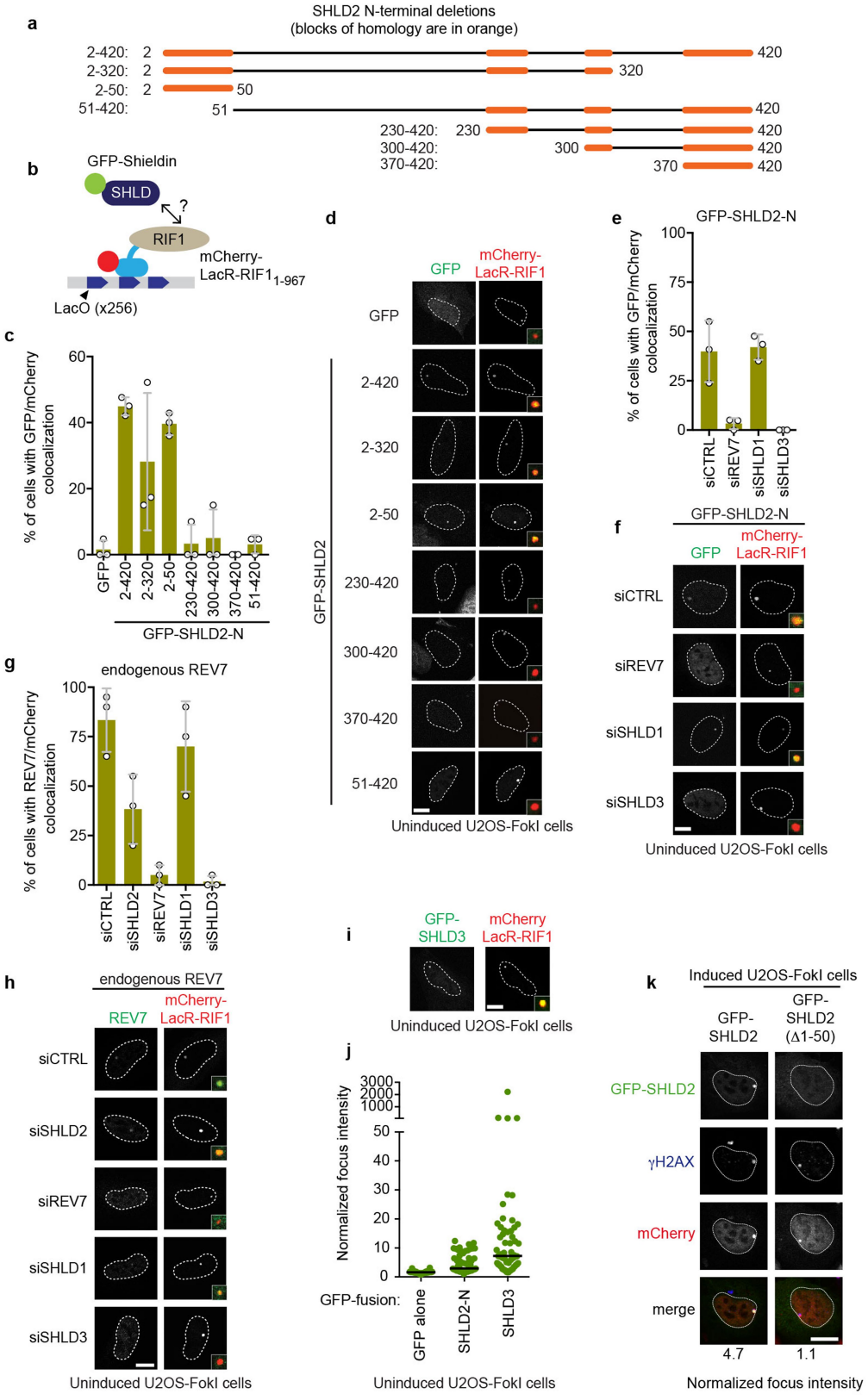
Extended Data Fig. 4 | Shieldin localizes to DSB sites. a, Representative micrographs of the experiments quantified in Fig. 3c. **b**, Representative micrographs of the experiments quantified in Fig. 3e. **c**, Quantification of mRNAs for *SHLD1*, *SHLD2* and *SHLD3*. RPE1 cells were transfected with siCTRL (non-targeting control siRNA) or siRNA targeting the indicated shieldin subunits. Forty-eight hours after transfection, mRNA was purified and reverse-transcribed before being assayed by quantitative real-time PCR. Data were normalized to the amount of *GAPDH* mRNA and expressed relative to the corresponding value for cells transfected with siCTRL. Data are presented as the mean ± s.d. ($n = 3$, independent siRNA transfections). **d**, Whole cell extracts from RPE1 wild-type cells transfected with the indicated siRNAs were processed for immunoblotting with the indicated antibodies. Tubulin is used as a loading control ($n = 1$ experiment; siRNA efficiency is also monitored by immunofluorescence). **e**, Quantification of 53BP1 and RIF1 recruitment to ionizing radiation-induced DSBs (1 h after irradiation with 10 Gy) following depletion of the indicated shieldin components. Data are represented as the mean ± s.d. ($n = 3$, independent siRNA transfections). **f**, Representative micrographs of the experiments quantified in **e**.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Epistasis between 53BP1 and shieldin factors. **a**, Quantification of RAD51 focus formation 3 h after irradiation (10 Gy) in RPE1 *BRCA1*^{KO} (left), *BRCA1*^{KO}*53BP1*^{KO} (middle) and *BRCA1*^{KO}*SHLD2*^{KO} (right) cells after viral transduction with the indicated sgRNAs (editing efficiency can be found in Supplementary Table 2) or empty vector (EV). Data are represented as the mean ± s.d. (for *BRCA1*^{KO}*53BP1*^{KO}, $n = 4$ biologically independent immunofluorescence experiments; for *BRCA1*^{KO} and *BRCA1*^{KO}*SHLD2*^{KO}, $n = 6$ biologically independent immunofluorescence experiments). *P* values were calculated using a two-tailed unpaired *t*-test. Left, *BRCA1*^{KO} EV versus sg53BP1-1 $P = 0.0002$; EV versus sgSHLD1-1 $P = 0.0043$; EV versus sgSHLD2-2 $P = 0.0348$; EV versus sgSHLD3-1 $P = 0.0180$; EV versus sgREV7-1 $P = 0.0012$). Middle, right: all comparisons to the EV condition were non-significant (NS). Values for *BRCA1*^{KO}*53BP1*^{KO} EV versus sg53BP1-1 $P = 0.2332$; EV versus sgSHLD1-1 $P = 0.4451$; EV versus sgSHLD2-2 $P = 0.9632$; EV versus sgSHLD3-1 $P = 0.1187$; EV versus sgREV7-1 $P = 0.0568$. Values for *BRCA1*^{KO}*SHLD2*^{KO}: EV versus sg53BP1-1 $P = 0.0550$; EV versus sgSHLD1-1 $P = 0.1864$; EV versus sgSHLD2-2 $P = 0.3568$; EV versus sgSHLD3-1 $P = 0.4641$; EV versus sgREV7-1 $P = 0.2888$. **b**, Talazoparib sensitivity of wild type or two independent *SHLD1*^{KO} SUM149PT-dox-Cas9 clones (A and D) virally transduced with an sgRNA targeting *53BP1* (sg53BP1) or a control non-targeting sgRNA (sgCtrl), following induction of Cas9. Data are presented as the mean ± s.d. ($n = 3$ biologically independent experiments).

© 2018 Springer Nature Limited. All rights reserved.

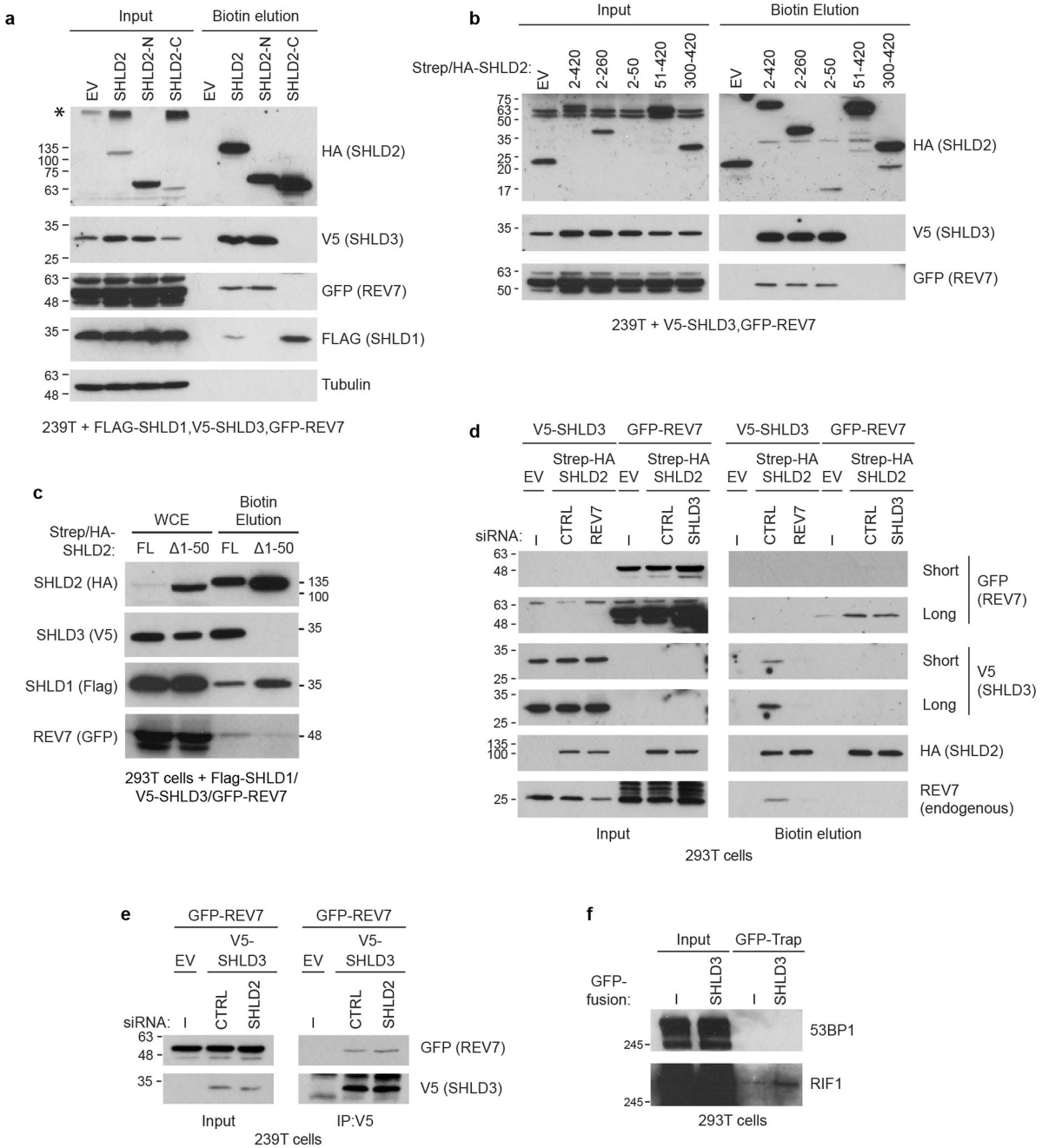


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

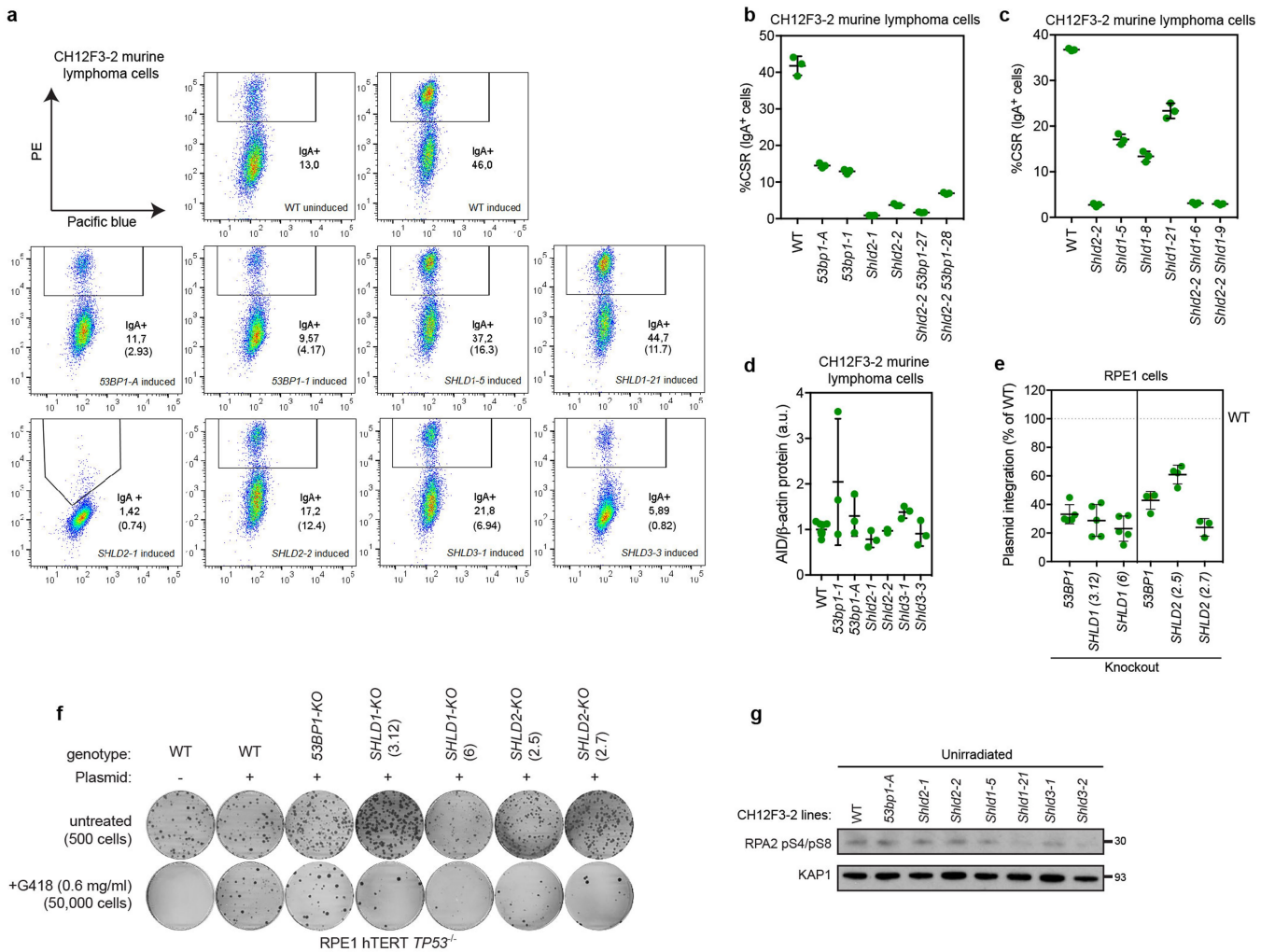
Extended Data Fig. 6 | The co-localization of shieldin with RIF1 on chromatin. a, Representation of the deletion mutants of SHLD2-N used in **c**, **d**. The orange shading indicates blocks of homology. **b**, Schematic of the LacR–RIF1 chromatin recruitment assay. **c**, Quantification of the experiment shown in **d**. Colocalization was considered positive when the average GFP intensity at the mCherry focus was threefold over background nuclear intensity. A minimum of 20 cells were imaged per biological replicate (circles); the bar represents the mean ± s.d. ($n = 3$). **d**, Representative images of the data quantified in **c**. The main focus is shown in inset ; scale bar, 10 μm. **e**–**h**, Quantification (**e**, **g**) and representative micrographs (**f**, **h**) of overexpressed GFP–SHLD2-N and mCherry–LacR–RIF1(1–967) co-transfected into uninduced U2OS–FokI cells along with siRNA against shieldin complex subunits after processing for mCherry and GFP (**e**, **f**) or mCherry and REV7 (**g**, **h**) immunofluorescence. Colocalization was considered positive when the average GFP or REV7 intensity at the mCherry focus was threefold over background nuclear intensity. A minimum of 20 cells were imaged per condition (circles); the bar represents the mean ± s.d. ($n = 3$ biologically independent experiments). **i**, Representative images of the data quantified in **j**. The main focus is shown in inset; scale bar, 10 μm. **j**, Quantification of GFP intensity at the mCherry–LacR–RIF1(1–967) focus, normalized to nuclear background. Each data point represents a cell transfected with the vector coding for the indicated GFP fusion. The line is at the median. The data are an aggregate of three independent experiments with a minimum of 20 cells counted (total cells counted: 62, 60 and 61 for GFP, GFP–SHLD2-C and GFP–SHLD3, respectively). **k**, mCherry–LacR–FokI colocalization with full-length or N-terminally truncated (Δ1–50) GFP–SHLD2. Mean normalized focus intensity is shown from a total of 59 (full-length SHLD2) or 56 (SHLD2 Δ1–50) cells counted ($n = 2$ biologically independent experiments).

© 2018 Springer Nature Limited. All rights reserved.



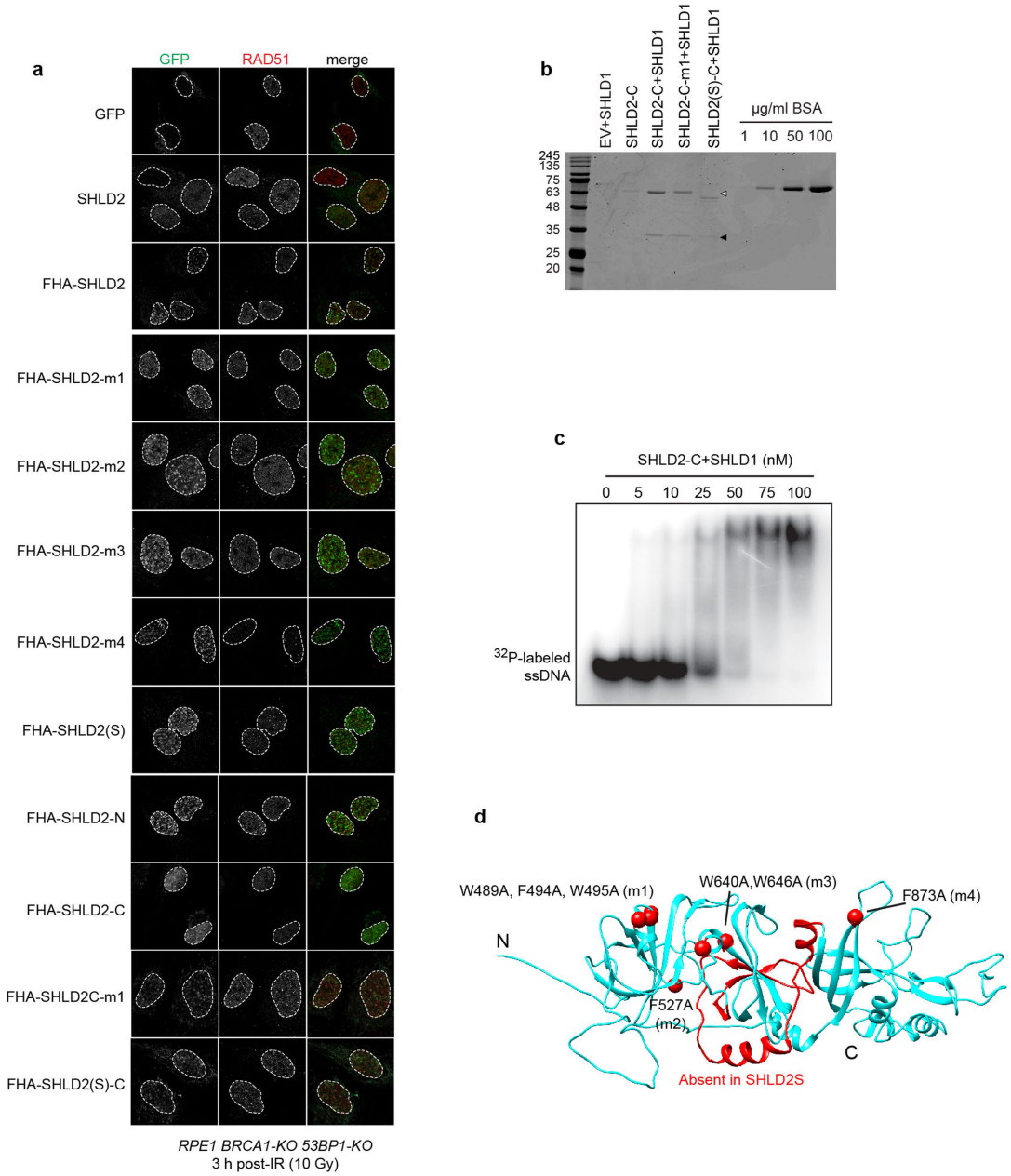
Extended Data Fig. 7 | Mapping the architecture of the shieldin complex. a, Streptavidin pulldown analysis determining which region of SHLD2 associates with the other shieldin subunits. WCEs of 293T cells transfected with an expression vectors for Flag–SHLD1, V5–SHLD3, GFP–REV7 and Strep/HA-tagged SHLD2, SHLD2-N (residues 2–420), SHLD2-C (residues 421–904) or empty Strep/HA vector (EV) were incubated with streptavidin resin and bound proteins were eluted with biotin. WCEs and elutions were analysed by SDS–PAGE and immunoblotting with the indicated antibodies. Tubulin was used as a loading control. Results are representative set of immunoblots from two independent experiments. Asterisk denotes a non-specific band. **b**, Mapping the SHLD3 and REV7 binding sites on the SHLD2 N terminus through streptavidin pulldowns with different SHLD2 constructs (detailed in Extended Data Fig. 6a) and immunoblotting. Results are a representative of a set of immunoblots from three independent experiments. **c**, Affinity purification of shieldin complex components using N-terminally truncated SHLD2 (Δ1–50) analysed by immunoblotting (representative of three independent experiments). **d**, Streptavidin pulldown analysis of SHLD2 association with REV7 and SHLD3. 293T cells were transfected with siRNAs and expression vectors for epitope-tagged shieldin components as indicated (EV, empty Strep/HA vector). WCEs were incubated with streptavidin resin and bound proteins were eluted with biotin. WCEs and elutions were analysed by SDS–PAGE and immunoblotted with the indicated antibodies. Short and long exposures are shown for GFP and V5 immunoblots ($n = 1$). **e**, Dependency of V5–SHLD3 co-immunoprecipitation with GFP–REV7. 293T cells were transfected with siRNAs and expression vectors for epitope-tagged REV7 and SHLD3 as indicated (EV, empty V5 vector). WCEs were incubated with anti-V5 antibody and protein G resin. Bound proteins were boiled in SDS sample buffer and analysed by immunoblotting with GFP and V5 antibodies ($n = 1$). **f**, Association between SHLD3 and RIF1. WCEs of 293T cells transfected with an expression vector for unfused GFP (−) or GFP–SHLD3 (SHLD3) were incubated with GFP-Trap resin. Bound proteins were boiled in SDS sample buffer and analysed by SDS–PAGE and immunoblotting against 53BP1 and RIF1. Results are representative of 2 SHLD3 immunoprecipitations, using SHLD3 fused to GFP (shown here) and V5 (shown in Fig. 3g) affinity tags.

© 2018 Springer Nature Limited. All rights reserved.



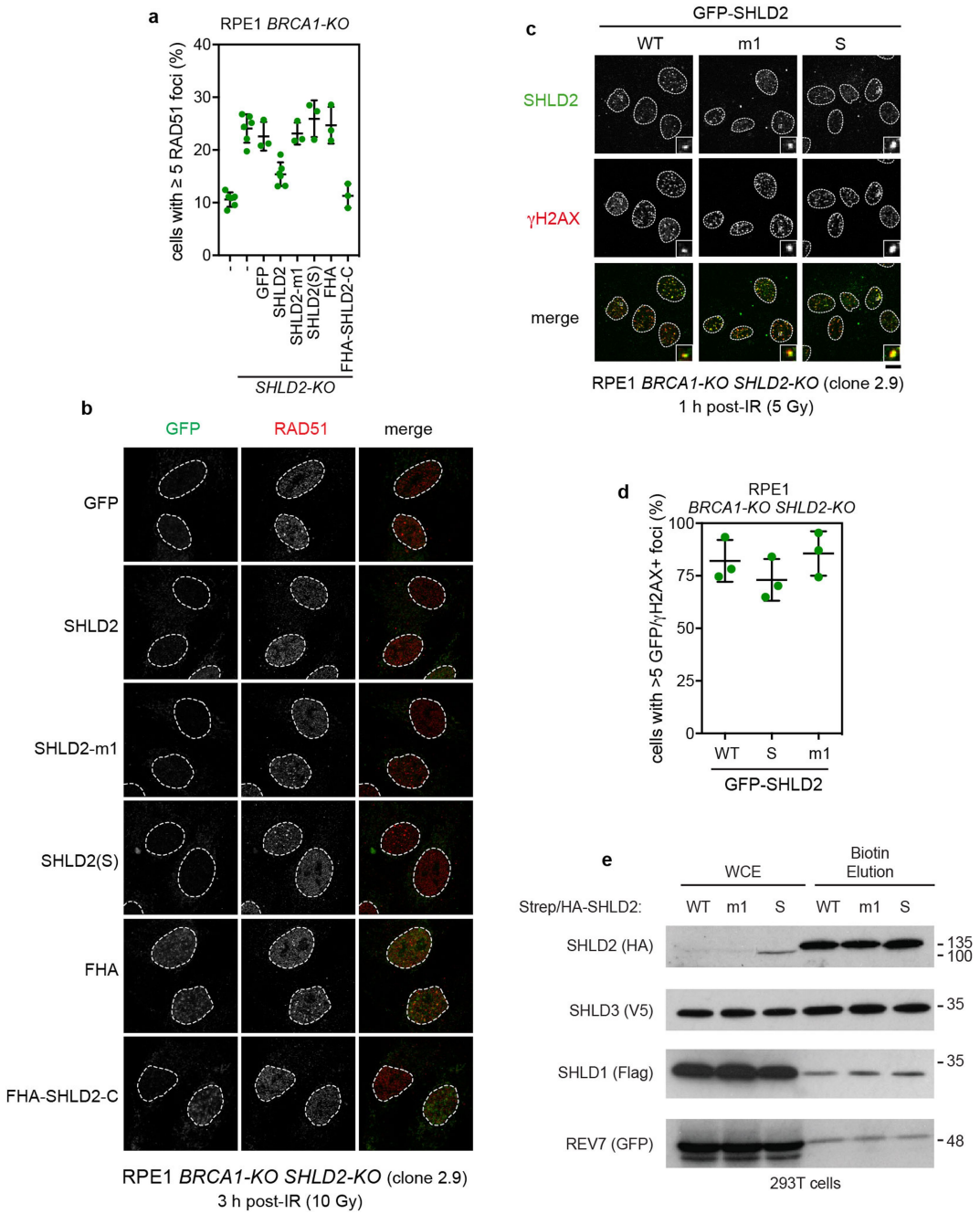
Extended Data Fig. 8 | Controls supporting the role of shieldin in promoting physiological NHEJ. a, Representative dot plots of the flow cytometry data obtained (of $n = 3$ biologically independent experiments) to assess class switching in Fig. 3h. Class switch recombination was determined as the percentage of IgA^+ cells following stimulation after subtracting the baseline percentage of IgA^+ cells in the indicated clones (values in parentheses). **b**, **c**, Epistasis analysis of shieldin and 53BP1 in class switch recombination. The percentage of class switching in CH12F3-2 wild type, single knockout or double knockout cells (as indicated) following stimulation is shown. Each data point represents a biological replicate; the line represents the mean ± s.d. ($n = 3$). Genomic editing efficiencies of the sgRNAs can be found in Supplementary Table 2. **d**, WCEs of the indicated CH12F2-3 clones were probed for AID and β-actin (loading control) by immunoblotting and were quantified by densitometry. Each data point represents a biological replicate; the line represents the mean ± s.d. ($n = 9$ for wild type, $n = 3$ for other samples). **e**, Random plasmid integration of linearized pcGFP-c1 conferring G418 resistance. Resistant colonies were quantified after 14 d. Bar represents the mean ± s.d. with wild-type cells set at 100% (left, $n = 5$; right, $n = 4$ except $SHLD2^{KO}$ (2.7) $n = 3$ biologically independent experiments). **f**, Representative images of the plasmid integration assays quantified in **e**. **g**, Un-irradiated CH12F3-2 clones were immunoblotted for RPA32 (also known as RPA2) phosphorylation (a representative set from $n = 3$ biological replicates; data relates to Fig. 3i).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | The role of DSB-targeted SHLD2 in the suppression of homologous recombination and the mapping of the SHLD2-C–SHLD1 complex binding to ssDNA. a, Representative micrographs of RPE1 *BRCA1*^{KO}*53BP1*^{KO} cells transduced with the indicated GFP-fusion proteins, pre-extracted, fixed and stained for RAD51 and GFP 3 h after ionizing radiation (10 Gy). Protein expression was induced for 24 h before exposure to ionizing radiation using 1 μg ml^{-1} doxycycline. Data relates to Fig. 4b. Note that owing to the pre-extraction required for visualization of RAD51 foci, the visualization of non-FHA-tagged SHLD2 is lost. **b**, SDS–PAGE analysis of purified SHLD2-C–SHLD1 complexes. Strep/HA–SHLD2(421–904)–Flag-SHLD1 complexes were purified from transiently transfected 293T cells. Concentrations of purified proteins were estimated by Coomassie staining and comparison to a standard curve of known BSA concentrations visualized by fluorescence at 700 nm. SHLD2-C m1 and SHLD2(S)-C denote SHLD2-C constructs carrying the OB-fold m1 mutation and the internal deletion (Δ655–723) corresponding to the naturally occurring splice variant of SHLD2, respectively. Open and filled arrowheads mark the bands corresponding to SHLD2-C and SHLD1, respectively. EV refers to empty Strep/HA vector. A representative stained gel from two independent experiments is shown. **c**, Representative image of the ^{32}P-labelled ssDNA EMSA with SHLD2-C–SHLD1 for K_d determination shown in Fig. 4e. **d**, Model of the SHLD2 OB-fold domains and the engineered mutations (red spheres, point mutations; red ribbons, splice variant deletion). Model relates to Fig. 4b, d.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | SHLD2 OB-folds are required for suppression of RAD51 focus formation induced by ionizing radiation. **a**, Quantification of RAD51 foci 3 h after 10 Gy irradiation in RPE1 *BRCA1*^{KO}*SHLD2*^{KO} cells complemented with the indicated GFP-tagged SHLD2 constructs via viral transduction. Protein expression was induced with 1 μg ml^{-1} doxycycline for 24 h before exposure to ionizing radiation. Each data point is a biological replicate; the bar represents the mean ± s.d. ($n = 6$ for *BRCA1*^{KO} untransduced cells, *BRCA1*^{KO}*SHLD2*^{KO} untransduced and GFP-SHLD2 cells, $n = 3$ for remaining cell lines, biologically independent experiments). **b**, Representative micrographs of the data shown in **a**. Note that owing to the pre-extraction required for visualization of RAD51 foci, the visualization of non-FHA tagged SHLD2 foci is lost. **c**, Representative micrographs of RPE1 *BRCA1*^{KO}*SHLD2*^{KO} cells virally transduced with vectors expressing GFP-tagged SHLD2 wild type or its OB-fold m1 mutant (m1), or short splice variant (S), 1 h after 5 Gy ionizing radiation. Scale bar, 10 μm. **d**, Quantification of the data shown in **c**. Each data point represents an independent biological replicate counting ≥ 100 cells. Data are represented as mean ± s.d. ($n = 3$). **e**, WCEs of 293T cells co-transfected with Strep/HA–SHLD2 wild type, Strep/HA–SHLD2 m1 or Strep/HA–SHLD2(S) mutants, and other shieldin subunits (Flag–SHLD1, V5–SHLD3, and GFP–REV7) were incubated with streptavidin resin and bound proteins were eluted with biotin. WCEs and eluted proteins were visualized by SDS–PAGE and immunoblotting with the indicated antibodies. Results shown are a representative set from two independent experiments.Source Data

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0362-1

53BP1 cooperation with the REV7–shieldin complex underpins DNA structure-specific NHEJ

Hind Ghezraoui[1,6], Catarina Oliveira[1,6], Jordan R. Becker[1], Kirstin Bilham[1], Daniela Moralli[2], Consuelo Anzilotti[3], Roman Fischer[4], Mukta Deobagkar-Lele[3], Maria Sanchiz-Calvo[1], Elena Fueyo-Marcos[1], Sarah Bonham[4], Benedikt M. Kessler[4], Sven Rottenberg[5], Richard J. Cornall[3], Catherine M. Green[2] & J. Ross Chapman[1]*

53BP1 governs a specialized, context-specific branch of the classical non-homologous end joining DNA double-strand break repair pathway. Mice lacking *53bp1* (also known as *Trp53bp1*) are immunodeficient owing to a complete loss of immunoglobulin class-switch recombination[1,2], and reduced fidelity of long-range V(D)J recombination[3]. The 53BP1-dependent pathway is also responsible for pathological joining events at dysfunctional telomeres[4], and its unrestricted activity in *Brca1*-deficient cellular and tumour models causes genomic instability and oncogenesis[5–7]. Cells that lack core non-homologous end joining proteins are profoundly radiosensitive[8], unlike 53BP1-deficient cells[9,10], which suggests that 53BP1 and its co-factors act on specific DNA substrates. Here we show that 53BP1 cooperates with its downstream effector protein REV7 to promote non-homologous end joining during class-switch recombination, but REV7 is not required for 53BP1-dependent V(D)J recombination. We identify shieldin—a four-subunit putative single-stranded DNA-binding complex comprising REV7, c20orf196 (SHLD1), FAM35A (SHLD2) and FLJ26957 (SHLD3)—as the factor that explains this specificity. Shieldin is essential for REV7-dependent DNA end-protection and non-homologous end joining during class-switch recombination, and supports toxic non-homologous end joining in *Brca1*-deficient cells, yet is dispensable for REV7-dependent interstrand cross-link repair. The 53BP1 pathway therefore comprises distinct double-strand break repair activities within chromatin and single-stranded DNA compartments, which explains both the immunological differences between *53bp1*- and *Rev7*- deficient mice and the context specificity of the pathway.

53BP1-dependent non-homologous end joining (NHEJ) requires the participation of several downstream factors[9,11,12], including REV7, which is a non-catalytic subunit of the translesion synthesis DNA polymerase-ζ. REV7 mediates genotoxic NHEJ events in *Brca1*-deficient mouse mammary tumour cells, and prevents end-resection at double-strand breaks (DSBs) associated with class-switch recombination (CSR)[13]. We generated a conditional *Rev7* knockout mouse (*Rev7*$^{f/f}$) in a C57BL/6 background, after germline *Rev7*-deletion resulted in embryonic lethality (Extended Data Table 1). *Rev7*$^{f/f}$ mice that possess the B cell *Mb1-cre* deleter allele[14] showed *Rev7* deletion at the start of the B cell lineage (Extended Data Fig. 1a), and normal B cell numbers with undetectable REV7 protein (Extended Data Fig. 1b). *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ mice expressed normal serum IgM but reduced titres of IgA and IgG, which is suggestive of CSR failure (Fig. 1a). Defective in vivo CSR was confirmed in immunization experiments in which serum concentrations of antigen-specific IgG1 in *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ mice were about 10-fold lower than in control *Rev7*$^{+/+}$*Mb1*$^{+/cre}$ mice, whereas IgM responses were comparable between groups (Fig. 1b). Likewise, *Rev7*-deletion severely compromized—by up to 90%—the production of class-switched B splenocytes upon stimulation in culture (Fig. 1c and Extended Data Fig. 1c), without affecting cell proliferation (Fig. 1d). Equivalent CSR frequencies in *Rev7*$^{f/f}$*Mb1*$^{+/cre}$, *53bp1*$^{-/-}$*Mb1*$^{+/cre}$ and *53bp1*$^{-/-}$*Rev7*$^{f/f}$*Mb1*$^{+/cre}$ double knockout cells furthermore confirmed REV7–53BP1 cooperation is essential for CSR (Fig. 1e).

However, notable differences in the absolute numbers of B lineage cells were detected between *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ and *53bp1*$^{-/-}$*Mb1*$^{+/cre}$ mice (Fig. 1f). Though *53bp1*-deficient mice showed 50% and 29% reductions in B220^{+} B lymphocytes in the bone marrow and spleen, respectively, these abnormalities were absent from *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ animals (Fig. 1f), leading us to question whether 53BP1-dependent DNA repair activities during B cell development require REV7. Detailed bone marrow analysis showed that B lymphocytes of *53bp1*$^{-/-}$*Mb1*$^{+/cre}$ mice became progressively depleted, with losses of approximately 70% of total lymphocytes by the late small pre-B and immature B cell stages (Hardy fractions D and E, respectively; Fig. 1g and Extended Data Fig. 2a). Losses were accompanied by increased apoptosis in bone marrow and follicular splenic B cell fractions (Extended Data Fig. 2b, c). By contrast, *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ mice showed normal B cell counts and apoptotic indices (Fig. 1g and Extended Data Fig. 2a–c), despite the complete absence of REV7 protein in B220^{+}CD43^{+} pro-B progenitors (Extended Data Fig. 2d). To exclude the possibility that developmental defects in *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ mice could be masked by compensatory changes, we generated chimaeric mice with mixed bone marrow cells. Equal mixes of CD45.1 wild-type and CD45.2 *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ whole bone marrow cells, or CD45.1 wild-type and CD45.2 *53bp1*$^{-/-}$*Mb1*$^{+/cre}$ whole bone marrow cells, were injected intravenously into lethally irradiated CD45.1 wild-type recipient mice (Extended Data Fig. 2e). Eight weeks after bone marrow transfer, the reconstitution of pro-, pre-, immature and mature bone marrow B cells derived from CD45.2 *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ cells did not differ from that of wild-type CD45.1 cells, similar to that in control recipient mice (data not shown). However, bone marrow B cells derived from CD45.2 *53bp1*$^{-/-}$*Mb1*$^{+/cre}$ mice only reconstituted about 30% of pro-B cell fractions, and were further outcompeted by wild-type CD45.1 bone marrow B cells at later stages, at which point they made up only about 15% of total immature and mature fractions (Extended Data Fig. 2e). These findings confirmed that B cell development is normal in the absence of REV7, which suggests that V(D)J recombination—which is essential for this process—does not require REV7.

To directly test whether 53BP1-dependent NHEJ during V(D)J recombination and CSR could be distinguished at the level of REV7 involvement, we monitored the stability of the *Igh* and *Ig*λ-light-chain (*Igl*) loci in stimulated B splenocytes from *Rev7*$^{f/f}$*Mb1*$^{+/cre}$, *53bp1*$^{-/-}$ and control (*Mb1*$^{+/cre}$) mice. Upon stimulation with anti-CD40 plus IL-4 in vitro, NHEJ-deficient splenic B cells accumulate *Igl* breaks owing to abortive RAG1- and RAG2-dependent secondary V–J recombination events, as well as *Igh* breaks associated with CSR[15]. Chromosome breakage at *Igh* and *Igl* was monitored by four-colour

[1]Genome Integrity Laboratory, Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Chromosome Dynamics, Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [3]MRC Human Immunology Unit, Weatherall Institute for Molecular Medicine, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [4]Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [5]Institute of Animal Pathology, Vetsuisse Faculty, University of Bern, Bern, Switzerland. [6]These authors contributed equally: Hind Ghezraoui, Catarina Oliveira. *e-mail: rchapman@well.ox.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

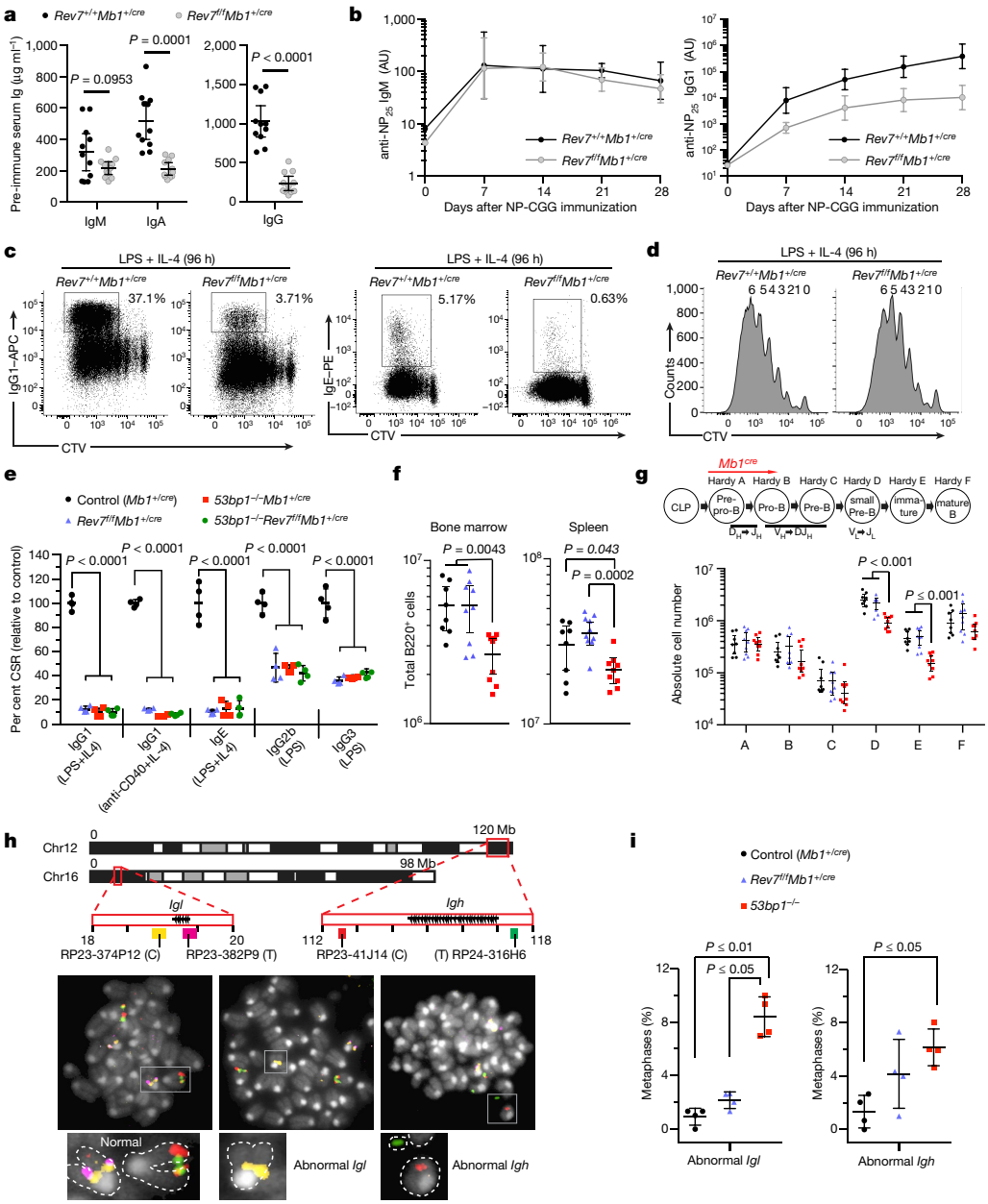


Fig. 1 | REV7 and 53BP1 cooperate during CSR, yet are functionally uncoupled in V(D)J recombination. a, Serum immunoglobulin in $Rev7^{+/+}Mb1^{+/cre}$ and $Rev7^{f/f}Mb1^{+/cre}$ mouse cohorts. $n = 11$ mice per genotype; *P* values, unpaired two-tailed *t*-test. Mean ± 95% confidence interval. **b**, NP-specific serum IgM (left) and IgG1 (right) at indicated times after NP-CGG immunization. AU, arbitrary units. Representative data, $n = 2$ independent experiments, each with 4 mice. Mean ± 95% confidence interval. **c**, Cell trace violet (CTV)-labelled splenic B cells were stimulated as indicated and stained for surface IgG1 or IgE on day 4. Representative data, $n > 6$ mice. **d**, CTV dilution in purified B cells cultured in the presence of LPS and IL-4 for 96 h. Representative data, $n > 6$ mice. **e**, Splenic B cells cultured with the indicated stimuli (96 h) and stained for surface IgG1, IgE, IgG2b or IgG3. $n = 4$ mice per genotype. CSR 100%, mean immunoglobulin isotype switch frequency of 2 control animals in each experiment. *P* values, two-way ANOVA with Tukey's correction. Mean ± 95% confidence interval. **f**, Absolute numbers of $B220^+$ B cells in the bone marrow (one femur plus one tibia) and spleen. $n = 9$ mice per genotype, except $Rev7^{+/+}Mb1^{+/cre}$ ($n = 8$). *P* values, unpaired two-tailed *t*-test, Mean ± 95% confidence interval. **g**, Absolute numbers of B cell precursors (Hardy[28] fraction A, $B220^+CD43^+BP\text{-}1^-CD24^-$; Hardy fraction B, $B220^+CD43^+BP\text{-}1^-CD24^+$; Hardy fraction C, $B220^+CD43^+BP\text{-}1^+CD24^+$; Hardy fraction D, $B220^+CD43^-IgM^-IgD^-$; Hardy fraction E, $B220^+CD43^-IgM^+IgD^-$; and Hardy fraction F, $B220^+CD43^-IgM^+IgD^+$) in the bone marrow (one femur and one tibia) from $Rev7^{+/+}Mb1^{+/cre}$ ($n = 8$), $Rev7^{f/f}Mb1^{+/cre}$ ($n = 9$) and $53bp1^{-/-}Mb1^{+/cre}$ ($n = 9$) mice. *P* values, unpaired two-tailed *t*-test. Mean ± 95% confidence interval. **h**, Top, schematic of the *Igh* and *Igl* loci and fluorescence in situ hybridization probes. Bottom, representative metaphase images showing normal and abnormal *Igh* and *Igl* loci. C, centromere-proximal; T, telomere-proximal. **i**, *Igh* and *Igl* locus breakage in splenic B cells of indicated mice upon stimulation (anti-CD40 + IL-4) for 96 h. $n = 4$ mice per genotype, between 98 and 151 metaphases were analysed from each mouse, except for one wild-type sample with only 45 metaphases; multiple comparisons, one-way ANOVA. Mean ± s.d.

fluorescence in situ hybridization on metaphase spreads, using probes that were positioned centromeric and telomeric to these loci[15]. Metaphases were classed as abnormal if either locus had bi-allelic centromeric BAC signals, with one allele lacking an adjacent signal from the telomeric BAC (Fig. 1h). As expected, $53bp1^{-/-}$ B cells had high levels of chromosomal abnormalities at both the *Igh* and *Igl* loci, unlike control cells for which locus breakage was rare (Fig. 1i). Three out of four $Rev7^{f/f}Mb1^{+/cre}$ mice exhibited breakage frequencies at the *Igh* locus equivalent to those in $53bp1^{-/-}$ mice. By contrast, *Igl* abnormalities in $Rev7^{f/f}Mb1^{+/cre}$ cells occurred at near-control frequencies,

© 2018 Springer Nature Limited. All rights reserved.

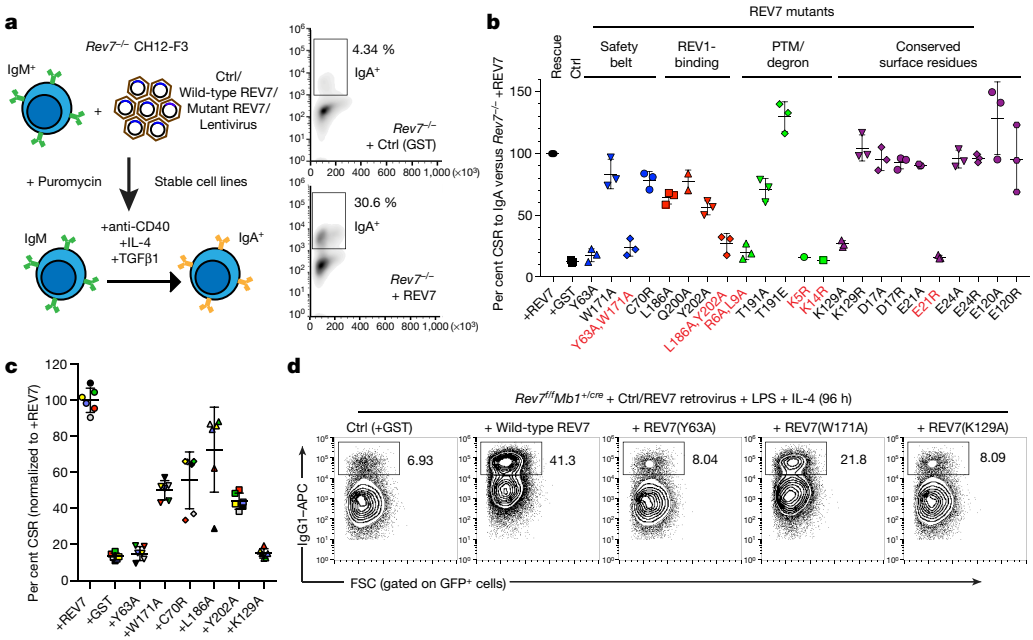


Fig. 2 | Mutational screening reveals REV7 interaction surfaces essential for CSR and resection inhibition. a, Schematic of the screen. *Rev7* single and combinatorial point mutant alleles were stably expressed in $Rev7^{-/-}$ CH12-F3 cells. IgM-to-IgA CSR was measured 40 h after stimulation with anti-CD40 antibody, IL-4 and TGFβ1. **b**, Quantification of IgM-to-IgA CSR in $Rev7^{-/-}$ CH12-F3 cell lines complemented as indicated. IgA switching efficiency normalized against wild-type REV7 complemented cells. Mutants indicated in red were excluded from further analysis owing to unstable REV7 protein expression. $n = 3$ independent experiments. Mean ± s.d. **c**, CSR in stimulated B cells collected from $Rev7^{f/f}Mb1^{+/cre}$ mice and infected with retrovirus expressing GST (control) or indicated REV7 protein. $n = 6$ mice per genotype. $IgG1^+$ events in infected cells as a proportion (%) of the mean $IgG1^+$ events in two wild-type REV7-complemented controls. $n = 3$ experiments (each with 2 mice per genotype). Mean ± s.d. **d**, Representative flow cytometry plots of data shown in **c**.

consistent with a selective requirement for 53BP1 but not REV7 in V(D)J recombination.

To determine the function of REV7 in CSR, we deleted *Rev7* in CH12-F3, a mouse B lymphoma cell line that undergoes efficient switching from IgM to IgA upon stimulation in vitro. CSR frequencies in $Rev7^{-/-}$ CH12-F3 cells were increased approximately 10-fold by reconstitution of REV7 expression compared to a GST control (Fig. 2a), and more than 20 REV7 single or combinatorial point mutants were screened for their ability to restore CSR (Fig. 2b and Extended Data Fig. 3a, b). Mutations that disrupt protein interactions[16,17], post-translational modifications or a destruction-box degron[18] were selected, and structure-led conservation analysis guided the generation of mutants within putative protein-interaction surfaces (summarized in Extended Data Table 2). Stable CSR-defective REV7 mutants were re-tested for their ability to support CSR to IgG1 upon complementation in stimulated B cells from $Rev7^{f/f}Mb1^{+/cre}$ mice (Fig. 2c, d). Two mutants, REV7(Y63A) and REV7(K129A), failed to rescue CSR to levels above those of control-complemented cells in both screens. The capacity of these mutants to prevent the hyper-resection of DSBs in the *Igh* switch region, which is a central function of the 53BP1–RIF1–REV7 axis during CSR[9,11,13,19], was therefore assessed by replication protein A (RPA)–single-stranded DNA (ssDNA) chromatin immunoprecipitation (ChIP) across *Igh* and control loci. As expected[13], DSB-associated donor ($S\mu$) and acceptor ($S\alpha$) switch loci, but not non-targeted *Igh* ($S\gamma1$) or control (*Rpp30*) loci, exhibited aberrant RPA–ssDNA enrichments in $Rev7^{-/-}$ cells (Extended Data Fig. 3c). RPA enrichments were suppressed in cell lines expressing wild-type REV7 but not the REV7(Y63A) or REV7(K129A) mutants, which confirms these mutations compromize resection inhibition (Extended Data Fig. 3c).

REV7 Tyr63 is one of two evolutionarily conserved residues that mediate interactions between the C-terminal 'safety-belt' domain and two conserved REV7-binding motifs in REV3L (RBM_1 and RBM_2; consensus PxxxpPSR)[17,20], each of which is essential for REV3L-dependent resistance to interstrand cross-link damage[20]. Alanine substitutions at Tyr63 or Trp171 in REV7 blocked interactions with the RBM_1 of REV3L, in a yeast two-hybrid (Y2H) assay (Extended Data Fig. 4a). As expected[21], $Rev7^{-/-}$ cells were sensitive to the interstrand cross-link inducer MMC, first accumulating in G2/M phase, and then progressing into cell death marked by an accumulation of sub-G1 events and profound chromosomal instability (Extended Data Fig. 4b–d). These defects were completely suppressed by complementation with wild-type REV7 but not the REV7(Y63A) or REV7(W171A) mutants, or the REV1-binding mutants REV7(L186A) or REV7(Y202A) (Extended Data Fig. 4b, e). By contrast, the CSR-deficient REV7(K129A) mutant completely restored wild-type responses to MMC, consistent with the unperturbed interaction between this mutant and REV3L (Extended Data Fig. 4a, e). This separation of function between REV7 interstrand cross-link repair and NHEJ activities implicates distinct REV7 complexes in NHEJ.

We therefore immunopurified control and Flag–HA–REV7 complexes from lysates of stably complemented $Rev7^{-/-}$ CH12-F3 cells, and analysed these by liquid chromatography–tandem mass spectrometry (LC–MS/MS). Aside from the known interactors REV3L and GTF2I[22], three uncharacterized proteins were highly enriched with REV7 (Fig. 3a and Table 1). The genes that encode these proteins, *c20orf196*, *FAM35A* and *FLJ26957* (also known as *CTC-534A2.2*)—renamed *SHLD1*, *SHLD2* and *SHLD3*, respectively—were cloned and the corresponding human protein was screened for interaction with wild-type and mutant REV7 by Y2H. In this assay, only SHLD3 showed direct interaction with REV7; this was abolished by REV7(Y63A) but was unaffected by REV7(W171A) (Fig. 3b and Table 1), which correlates perfectly with the requirement for Tyr63 but not Trp171 in NHEJ. SHLD3 comprises two N-terminal REV3L-like RBM motifs, and predicted structural folds that resemble the mRNA cap-binding domain of the translational elongation initiation factor EIF4E (Fig. 3c; folds recognized by Phyre2[23]). REV7–SHLD3 interactions rely predominantly on Pro53 and Pro58 in RBM_2 (Fig. 3c), which is a motif that resembles RBM_1 in REV3L (Extended Data Fig. 5a).

The CSR defects exhibited by REV7(K129A) could not be explained by a failure to interact with SHLD3 (Fig. 3b). Lys129 is a highly

© 2018 Springer Nature Limited. All rights reserved.

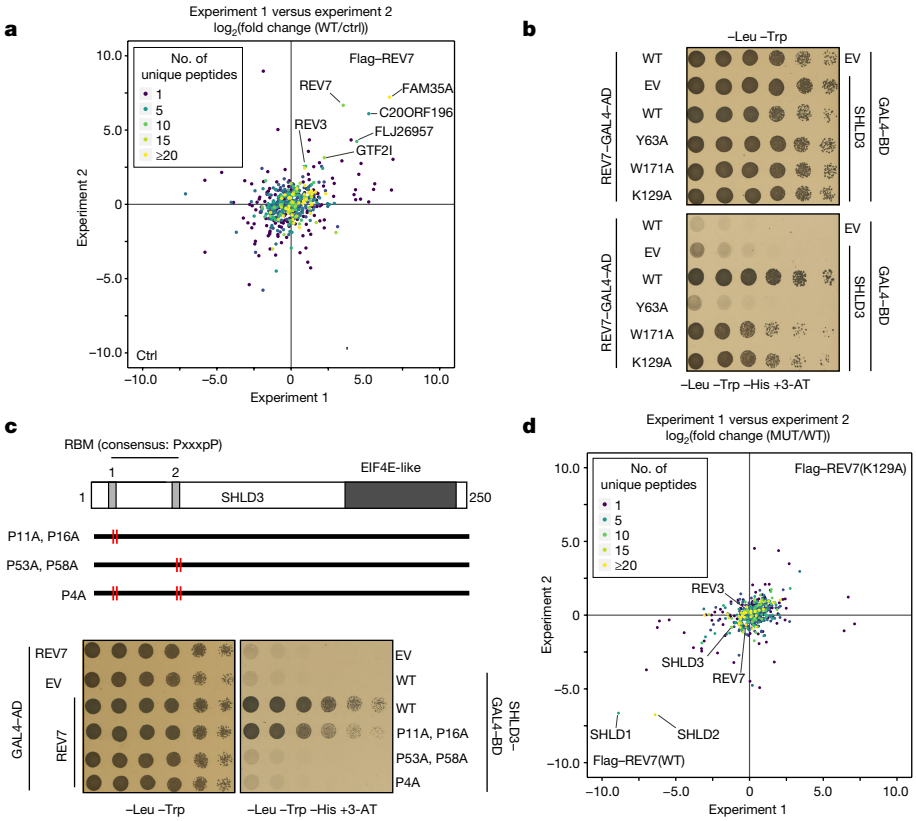


Fig. 3 | REV7 interacts with SHLD3 and SHLD1–SHLD2 via distinct surfaces. a, B cell REV7 interactome as defined by LC–MS/MS and LFQ. Scatter plot depicts $\log_2$ fold-enrichment of indicated Flag–immunocomplexes across 2 independent experiments. **b**, Interactions between indicated control (AD or BD empty vector (EV)), SHLD3 and indicated REV7 proteins by Y2H. Fivefold dilutions, representative data ($n = 4$). **c**, Top, schematic of SHLD3, indicating sequence features and positions of alanine substitutions in each mutant protein. Bottom, interactions between REV7 and indicated SHLD3 proteins by Y2H. Representative data ($n = 4$). **d**, LC–MS/MS and LFQ analysis of Flag–HA–REV7 and Flag–HA–REV7(K129A) interactomes. The scatter plot depicts $\log_2$ fold-enrichment of interacting proteins across 2 independent experiments. WT, wild type.

conserved residue on an uncharacterized structural surface that was selected upon analysis of the REV7 crystal structure[16] (Extended Data Fig. 5b). A mutant containing a conservative arginine substitution of Lys129 completely restored wild-type CSR frequencies (Fig. 2b and Extended Data Fig. 3b), which eliminated consideration of post-translational modification. Therefore, we compared the compositions of the Flag–HA–REV7(K129A) and wild-type Flag–HA–REV7 complexes using LC–MS/MS and label-free quantification (LFQ). Consistent with Y2H data, Flag–HA–REV7(K129A) complexes retained equivalent levels of REV3L and SHLD3 but were devoid of SHLD1–SHLD2 (Fig. 3d and Table 1), implicating SHLD1 and SHLD2 in NHEJ.

Isogenic knockout CH12-F3 clones were therefore generated for *Shld1*, *Shld2* and *Shld3*. No changes in REV7 or 53BP1 expression were identified in these knockout cell lines (Extended Data Fig. 6a), which suggests a direct role for a complex formed by REV7 and all three SHLD proteins in NHEJ. Normal proliferation but diminished CSR was detected in all clones (Fig. 4a, b and Extended Data Fig. 6b). CSR was restored in $Shld3^{-/-}$ and $Shld2^{-/-}$ cells by integration of *Shld3* and *Shld2* transgenes, respectively (Fig. 4c and Extended Data Fig. 6c), and mutant *Shld3* transgenes revealed that RBM_1 and RBM_2 support CSR redundantly (Fig. 4c. and Extended Data Fig. 6d). *Shld3* deletion did not augment CSR defects in $Rev7^{-/-}$ CH12-F3 cells (Extended Data Fig. 6e), confirming epistasis between *Shld3* and *Rev7* in NHEJ. These data indicate REV7 forms the linchpin in a four-subunit protein complex, the integrity of which is essential for 53BP1-dependent CSR. The C-terminal safety-belt of REV7 mediates interaction with SHLD3, and residues centred around Lys129 mediate binding to SHLD1–SHLD2 heterodimers. Structural modelling of SHLD2 revealed a triple-tandem OB-fold architecture with structural homology to the core ssDNA-binding OB-folds 2–4 in fungal RPA70[24] (Extended Data Fig. 5c, d). We and others[25] have named this putative ssDNA-binding complex 'shieldin' and its subunits as SHLD1, SHLD2 and SHLD3, by analogy with the telomere end-capping complex shelterin, the ssDNA and double-stranded DNA (dsDNA) binding proteins of which cooperate in chromosome end protection[26].

RPA–ssDNA ChIP experiments confirmed the contribution of shieldin to DNA-end protection: RPA–ssDNA complexes were undetectable at donor ($S\mu$) and acceptor ($S\alpha$) loci in stimulated cultures of wild-type and SHLD3-complemented $Shld3^{-/-}$ CH12-F3 cells, yet highly enriched in $Rev7^{-/-}$, $Shld2^{-/-}$ and $Shld3^{-/-}$ cells (Fig. 4d and Extended Data Fig. 6f, g). Thus, REV7–shieldin inhibits DSB resection during NHEJ. However, $Shld3^{-/-}$ cell lines were indistinguishable from control upon MMC treatment, consistent with REV7

Table 1 | Label-free quantification of LC–MS/MS results for indicated interacting proteins

	Flag–HA–REV7 (wild type)				
Protein symbol	Number of unique peptides	Coverage (%)	LFQ[a]	LFQ[b]	Y2H interaction with REV7
REV7 (MAD2L2)	9	42	5.1	−0.5	NA
c20orf196 (SHLD1)	5	34	5.7	−7.8	–
FAM35A (SHLD2)	15	26	6.9	−6.6	–
FLJ26957 (SHLD3)	8	27	4.3	−1.01	+
REV3L	8	3	1.8	0.02	+
GTF2I	10	12	2.7	0.95	ND

Interacting proteins were determined as in Fig. 3a, d. NA, not applicable; ND, not determined. $n = 2$ independent experiments.
[a]Average $\log_2$(fold change(wild type/control)), $n = 2$.
[b]Average $\log_2$(fold change(K129A/wild type)), $n = 2$.

© 2018 Springer Nature Limited. All rights reserved.

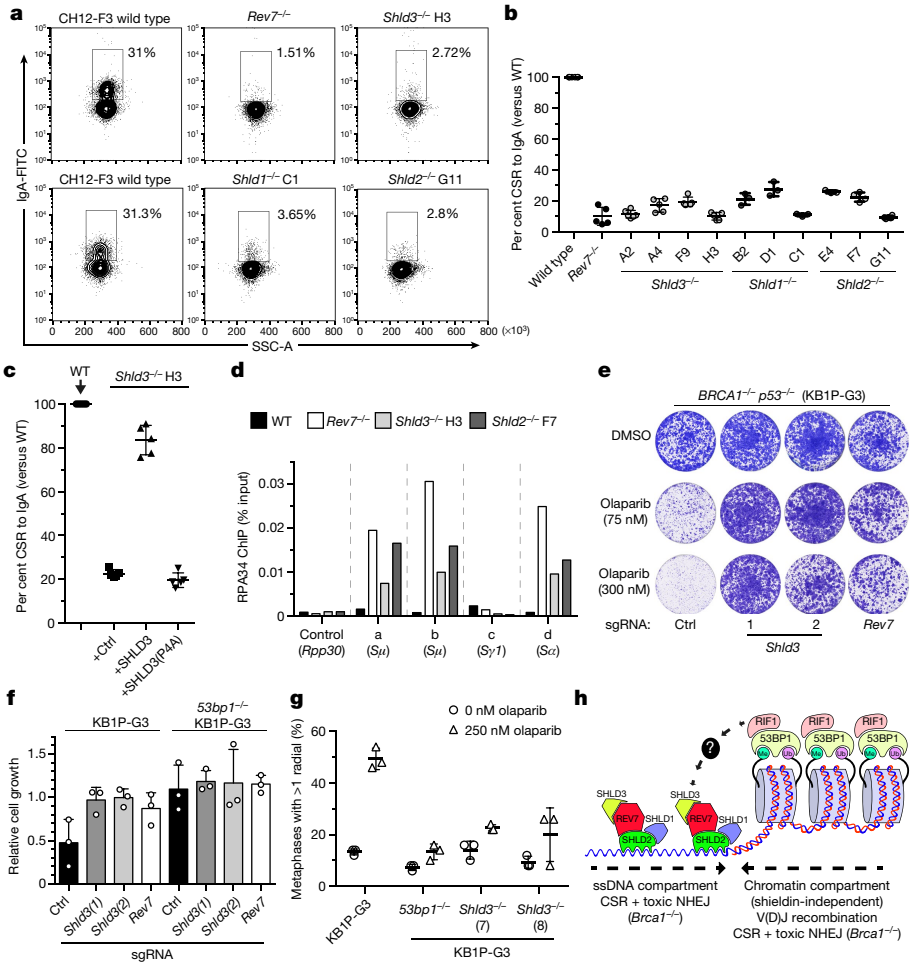


Fig. 4 | The REV7–shieldin complex mediates 53BP1-dependent NHEJ. **a**, Representative flow cytometry plots for IgM-to-IgA CSR in indicated mutant CH12-F3 cell lines. Representative data, $n > 3$ independent experiments. **b**, IgM-to-IgA CSR frequencies in knockout CH12-F3 cell lines. $n = 5$ (*Shld3*$^{-/-}$), $n = 4$ (*Shld1*$^{-/-}$) and $n = 3$ (*Shld2*$^{-/-}$) independent experiments. Mean ± s.d. **c**, IgM-to-IgA CSR in indicated CH12-F3 lines stably transduced with either control (GST), wild-type or mutant-SHLD3-expressing transgenes. Data normalized to CSR frequencies of wild-type CH12-F3 cells. $n = 5$ independent experiments. Mean ± s.d. **d**, RPA–ssDNA ChIP with indicated CH12-F3 lines stimulated with anti-CD40 antibody, IL-4 and TGFβ1 (30 h). Representative data, $n = 2$ independent experiments. Bars indicate mean. **e**, Clonogenic assay using KB1P-G3 cells following deletion of indicated genes. sgRNA, gene targeted by CRISPR–Cas9 guide RNA. **f**, Quantification of data in **e** and Extended Data Fig. 6k. $n = 3$ independent experiments. Mean ± s.d. **g**, Metaphases prepared from indicated cell lines following 24-h control or olaparib treatment were analysed for the presence of radial chromosomes. $n = 3$ independent experiments (each 50 metaphases per condition). Mean ± s.d. **h**, Proposed model of shieldin–53BP1 cooperation during NHEJ. The stabilization of ssDNA overhangs or ssDNA–dsDNA junctions at DSB sites by REV7–SHLD1–SHLD2–SHLD3 complexes promotes DSB resolution activities during NHEJ. The absence of lymphocyte development defects in *Rif1*$^{-/-}$ mice[9] suggests RIF1 could link 53BP1 anti-resection activities in chromatin to shieldin activities in ssDNA compartments (arrow).

separation-of-function in interstrand cross-link repair and NHEJ (Extended Data Fig. 6h).

We next investigated the contribution of shieldin to 53BP1-dependent toxic NHEJ in *Brca1*-deficient cells. CRISPR–Cas9-mediated mutagenesis of *Shld3* in the *Brca1*$^{-/-}$*p53*$^{-/-}$ mouse mammary tumour KB1P-G3 cell line[13] was strongly selected for in the presence of the poly(ADP-ribose) polymerase inhibitor (PARPi) olaparib (Extended Data Fig. 6i). Moreover, PARPi resistance in *Shld3*$^{-/-}$ KBP1-G3 cells was equivalent to *Rev7*$^{-/-}$ KBP1-G3 controls (Fig. 4e, f). There was neither selection for mutagenesis of *Shld3* or *Rev7* loci in olaparib-treated *53bp1*$^{-/-}$ KB1P-G3 cells, nor increased resistance in *Shld3*$^{-/-}$*53bp1*$^{-/-}$ cells (Fig. 4f and Extended Data Fig. 6j, k). Similar reductions in olaparib-induced radial chromosomes in both *Shld3*$^{-/-}$ and *53bp1*$^{-/-}$ KB1P-G3 lines confirmed toxic NHEJ in *Brca1*-deficient cells is dependent on shieldin (Fig. 4g). Shieldin disruption therefore provides a possible route to PARPi resistance in cancer.

This study reveals a requirement for a putative ssDNA-binding protein complex, shieldin, in the 53BP1 pathway, indicating that multiple DNA binding activities cooperate during DNA structure-specific NHEJ (Fig. 4h). The coupling of 53BP1-dependent anti-resection activities within chromatin to shieldin-dependent stabilization of ssDNA-tailed ends may permit the conversion of ssDNA-tailed substrates, such as those generated during CSR, into DSBs amenable to NHEJ. This segregation of distinct yet cooperative activities at DSB sites can explain the developmental differences between 53BP1- and REV7-deficient lymphocytes: ssDNA-tails of AID-dependent DSBs induced during CSR require these specialized activities to orchestrate their joining by NHEJ[27], whereas the absence of ssDNA at RAG-induced DSBs produced during V(D)J recombination could preclude the need for shieldin. ssDNA-tailed DSBs additionally exist at uncapped telomeres, and potentially at collapsed replication forks, and thus a requirement for shieldin during the repair of these structures can explain the physiological and pathophysiological specificities of the 53BP1 system.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0362-1.

© 2018 Springer Nature Limited. All rights reserved.

Received: 26 January 2018; Accepted: 13 June 2018;
Published online 25 July 2018.

1. Ward, I. M. et al. 53BP1 is required for class switch recombination. *J. Cell Biol.* **165**, 459–464 (2004).
2. Manis, J. P. et al. 53BP1 links DNA damage-response pathways to immunoglobulin heavy chain class-switch recombination. *Nat. Immunol.* **5**, 481–487 (2004).
3. Difilippantonio, S. et al. 53BP1 facilitates long-range DNA end-joining during V(D)J recombination. *Nature* **456**, 529–533 (2008).
4. Dimitrova, N., Chen, Y.-C. M., Spector, D. L. & de Lange, T. 53BP1 promotes non-homologous end joining of telomeres by increasing chromatin mobility. *Nature* **456**, 524–528 (2008).
5. Cao, L. et al. A selective requirement for 53BP1 in the biological response to genomic instability induced by Brca1 deficiency. *Mol. Cell* **35**, 534–541 (2009).
6. Bunting, S. F. et al. 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. *Cell* **141**, 243–254 (2010).
7. Bouwman, P. et al. 53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers. *Nat. Struct. Mol. Biol.* **17**, 688–695 (2010).
8. Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat. Rev. Mol. Cell Biol.* **4**, 712–720 (2003).
9. Chapman, J. R. et al. RIF1 is essential for 53BP1-dependent nonhomologous end joining and suppression of DNA double-strand break resection. *Mol. Cell* **49**, 858–871 (2013).
10. Cuella-Martin, R. et al. 53BP1 integrates DNA repair and p53-dependent cell fate decisions via distinct mechanisms. *Mol. Cell* **64**, 51–64 (2016).
11. Di Virgilio, M. et al. Rif1 prevents resection of DNA breaks and promotes immunoglobulin class switching. *Science* **339**, 711–715 (2013).
12. Escribano-Díaz, C. et al. A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. *Mol. Cell* **49**, 872–883 (2013).
13. Xu, G. et al. REV7 counteracts DNA double-strand break resection and affects PARP inhibition. *Nature* **521**, 541–544 (2015).
14. Hobeika, E. et al. Testing gene function early in the B cell lineage in mb1-cre mice. *Proc. Natl Acad. Sci. USA* **103**, 13789–13794 (2006).
15. Wang, J. H. et al. Mechanisms promoting translocations in editing and switching peripheral B cells. *Nature* **460**, 231–236 (2009).
16. Wojtaszek, J. et al. Structural basis of Rev1-mediated assembly of a quaternary vertebrate translesion polymerase complex consisting of Rev1, heterodimeric polymerase (Pol) ζ, and Pol κ. *J. Biol. Chem.* **287**, 33836–33846 (2012).
17. Hara, K. et al. Crystal structure of human REV7 in complex with a human REV3 fragment and structural implication of the interaction between DNA polymerase ζ and REV1. *J. Biol. Chem.* **285**, 12299–12307 (2010).
18. Listovsky, T. & Sale, J. E. Sequestration of CDH1 by MAD2L2 prevents premature APC/C activation prior to anaphase onset. *J. Cell Biol.* **203**, 87–100 (2013).
19. Hakim, O. et al. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. *Nature* **484**, 69–74 (2012).
20. Tomida, J. et al. REV7 is essential for DNA damage tolerance via two REV3L binding sites in mammalian DNA polymerase ζ. *Nucleic Acids Res.* **43**, 1000–1011 (2015).
21. Bluteau, D. et al. Biallelic inactivation of REV7 is associated with Fanconi anemia. *J. Clin. Invest.* **126**, 3580–3584 (2016).
22. Fattah, F. J. et al. The transcription factor TFII-I promotes DNA translesion synthesis and genomic stability. *PLoS Genet.* **10**, e1004419 (2014).
23. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **10**, 845–858 (2015).
24. Fan, J. & Pavletich, N. P. Structure and conformational change of a replication protein A heterotrimer bound to ssDNA. *Genes Dev.* **26**, 2337–2347 (2012).
25. Noordermeer, S. M. et al. The shieldin complex mediates 53BP1-dependent DNA repair. *Nature* https://doi.org/10.1038/s41586-018-0340-7 (2018).
26. Palm, W. & de Lange, T. How shelterin protects mammalian telomeres. *Annu. Rev. Genet.* **42**, 301–334 (2008).
27. Stavnezer, J. & Schrader, C. E. IgH chain class switch recombination: mechanism and regulation. *J. Immunol.* **193**, 5370–5378 (2014).
28. Hardy, R. R., Carmack, C. E., Shinton, S. A., Kemp, J. D. & Hayakawa, K. Resolution and characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow. *J. Exp. Med.* **173**, 1213–1225 (1991).

Acknowledgements We thank members of the Chapman, Green, Cornall and Rottenberg laboratories for discussions; T. Humphrey for comments on the manuscript; D. Adams for the *Rev7*^{tm1a} mouse strain; S. Boulton for *53bp1*$^{-/-}$ mice; J. Grimes and R. Nolan for assistance with protein modelling and statistics; and B. Davies, M. Barazas, B. Reina-San-Martin, B. Deplancke, L. Vasilieva, A. Nussenzweig and E. Callen-Moreau for reagents and advice. H.G. thanks K. Ghezraoui for her unwavering support and encouragement. This project was funded by Medical Research Council (MRC) Grant (MR/M009971/1) and Cancer Research UK Career Development Fellowship (C52690/A19270) awarded to J.R.C. C.A. and M.D.-L. are funded by the MRC, R.J.C. is a Principal Investigator of the MRC Human Immunology Unit. R.F. and B.M.K. are supported by the Kennedy Trust and the John Fell Fund. M.S.-C. and E.M.-F. were supported by ERASMUS+ fellowships. The Wellcome Centre for Human Genetics is supported by Wellcome grant (090532/Z/09/Z).

Author contributions H.G. and C.O. designed and performed the majority of the experiments, and analysed the data with assistance from K.B., C.A., M.D.-L., M.S.-C., E.F.-M. and R.J.C. J.R.B. designed and performed the KB1P-G3 experiments, and analysed the data. K.B. performed and analysed the immunization and ELISA experiments, established the mixed chimaeras and coordinated animal experiments. D.M. and C.M.G. performed and analysed all fluorescence in situ hybridization and metaphase chromosome experiments. R.F. ran and analysed the LC–MS/MS experiments with assistance from S.B. and supervision from B.M.K. This study was initiated in collaboration with S.R. J.R.C. conceived and supervised the study, performed experiments, analysed the data, prepared the figures and wrote the manuscript with editorial contributions from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0362-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0362-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.R.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Mice. All mice used for this study were generated on, or backcrossed onto a C57BL/6 background (>10 generations). Sperm from mice containing the *Mad2l2*$^{tm1a(EUCOMM)Wtsi}$ knockout-first conditional allele for *Rev7* (MGI:4432091, a gift from D. Adams, Wellcome Sanger Institute, Hinxton), was used to re-derive *Mad2l2*$^{tm1a(EUCOMM)Wtsi/+}$ mice, which were then bred with constitutive *Flp*-deleter mice (Tg(*ACTB-Flpe*)9205Dym; Jax stock 005703) to generate mice with the *Mad2l2*^{tm1c} conditional allele. Experimental *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ (*Mad2l2*$^{tm1c/tm1c}$) and related control mice were generated by intercrossing with mice containing the B cell lineage *Mb1-cre* deleter strain (*Cd79a*$^{tm1(cre)Reth}$; MGI:3687451[14]). All experiments involved age-matched 8–16-week-old animals on an inbred C57BL/6 background. *53bp1*$^{-/-}$ mice (MGI:2654201; also C57BL/6) were generated and described elsewhere[29]. For bone marrow mixed chimaera experiments, wild-type C57BL/6-Ly5.1 mice (hereafter known as CD45.1; Jax stock 002014) were irradiated with two doses of 4.5 Gy spaced 3 h apart, and subsequently injected with 5–10 × 10^6 bone marrow cells (approximately 50:50 mixture) of wild-type CD45.1$^+$ bone marrow and either wild-type *Mb1*$^{+/cre}$, *Rev7*$^{f/f}$*Mb1*$^{+/cre}$ or *53bp1*$^{-/-}$*Mb1*$^{+/cre}$ (CD45.2$^+$) bone marrow. Baytril 10% antibiotic solution (Vet Services, University of Oxford) was added to drinking water (1.5 ml to 250 ml water) for 3 weeks after irradiation. Mice were allowed to reconstitute for 7–8 weeks before analysis. Sample sizes were determined by power calculations. We were not pursuing lower penetrance phenotypes, thus statistically significant data could typically be obtained with small group sizes (typically, 4–10 mice). Randomization of samples was only undertaken during the scoring of chromosomal aberration during metaphase analyses. Mice of a certain genotype were selected based on a unique mouse ID number that does not indicate mouse genotype, thus phenotype–genotype relationships were determined only at the data analysis stage. All experiments were approved by the University of Oxford Ethical Review Committee and performed under a UK Home Office Licence in compliance with animal use guidelines and ethical regulations.

Immunizations. *Mb1*$^{+/cre}$, *Rev7*$^{f/f}$*Mb1*$^{+/cre}$, *or Rev7*$^{f/-}$*Mb1*$^{+/cre}$ mice were immunized intraperitoneally with 50 mg of NP-CGG (Santa Cruz Biotechnologies) emulsioned in Imject Alum adjuvant (Pierce, Thermo Fisher Scientific). Blood samples were collected from the tail vein at 0, 7, 14, 21 and 28 days after immunization.

Enzyme-linked immunosorbent assays. Enzyme-linked immunosorbent assays (ELISAs) were used to quantify the production of NP-specific antibodies in mice serum. Ninety-six-well plates were coated with 1 μg/ml NP-BSA (Biosearch Technologies) in bicarbonate buffer, blocked with 5% milk in PBS and incubated with serial dilutions of serum collected at different time points from immunized mice. Plates were probed using alkaline phosphatase-coupled antibodies against mouse IgM and IgG1 (Southern Biotech). Phosphatase substrate (Sigma) was used for detection and optical density measured at 405 nm. For IgG1, pooled blood from post-immunization wild-type mice was used as a standard and serially diluted into a standard curve. The first dilution was established as 1,000 arbitrary units. For IgM pooled blood from day 7 was used as a standard. Immunoglobulin concentrations in mouse serum or culture supernatants were determined by sandwich ELISA. Total IgG, IgM and IgA was measured with mouse IgG, IgM and IgA ELISA kits, respectively (Bethyl Laboratories), according to the manufacturer's instructions. Mouse serum with known immunoglobulin concentrations of each immunoglobulin was used as a standard.

B lymphocyte analysis and flow cytometry. Cell suspensions from bone marrow (one femur and one tibia) and spleen were counted on a haemocytometer and stained with anti-mouse antibodies against the following antigens as appropriate (all from BioLegend unless otherwise stated) in FACS buffer (PBS with 2% BSA and 0.025% sodium azide): IgD (1:500, 405716 or Thermo Fisher 12-5993-82; 11-26c.2a), IgG1 (1:200, 406606; RMG1-1), IgM (1:500, 406506 or 406514; RMM-1), B220 (1:500, 103232, 103244, 103206, or 103212; RA3-6B2), BP-1 (1:200, 108308; 6C3), CD19 (1:500, 115534 or 115520; 6D5), CD24 (1:1000, 101827 or BD Pharmingen 562563; M1/69), CD93 (1:200, 136510; AA4.1), CD23 (1:200, BD Pharmingen 553139; B3B4), CD43 (1:200, BD Pharmingen 562865; S7, or Thermo Fisher 11-0431-85; eBioR2/60), and CD21 (1:500, BD Biosciences 563176; 7G6). Mouse BD Fc Block (1:500, BD Pharmingen 553141) was added to block non-specific binding and live/dead cells were discriminated after staining with Zombie Aqua viability dye (1:200, 423102) or Zombie NIR viability dye (1:500, 423105). Bone marrow cells from mixed chimaeras were also stained against CD45.1 (1:500, 110730; A20) and CD45.2 (1:500, 109818; 104). Data were acquired on a FACSCanto (BD Biosciences), SH800 (Sony), or MoFlo Legacy (Beckman Coulter) and were analysed with FlowJo Software v10 (Tree Star). Gating strategies for bone marrow (left panel) and spleen (right panel) are shown in Supplementary Fig. 2; further Hardy fraction gating strategies are shown in Extended Data Fig. 2a.

Ex vivo B splenocyte culture, stimulation and flow cytometry. B cells were purified from red blood cell-lysed single-cell suspensions of mouse spleens by magnetic negative selection using a B Cell Isolation Kit (Miltenyi Biotec, 130-090-862). B cells (3 × 10^5 per well in a 12-well plate) were cultured in RPMI supplemented with 10% FCS, 100 U/ml penicillin, 100 ng/ml streptomycin, 2 mM L-glutamine, 1× MEM nonessential amino acids, 1 mM sodium pyruvate and 50 μM β-mercaptoethanol. B cells were stimulated with 5 μg/ml LPS (Sigma, L7770-1MG), 10 ng/ml mouse recombinant IL-4 (Peprotech, 214-14-20), and agonist anti-CD40 antibody (0.5 μg/ml; Miltenyi Biotec; FGK45.5). Cultures were grown at 37 °C with 5% CO_2 under ambient oxygen conditions. Four days after seeding, stimulated B cells were analysed using a FACSCanto; analysis was performed using FlowJo. Cells were resuspended in FACS buffer, blocked with Mouse BD Fc Block, and immunostained with biotinylated antibodies as follows: anti-mouse IgG1 (1:100, BD Pharmingen 553441; A85-1), anti-mouse IgG2b (1:100, BioLegend 406704; RMG2b-1), anti-mouse IgG3 (1:100, BD Pharmingen 553401; R40-82) and Streptavidin APC (1:500, Thermo Fisher 17-4317-82). Cells expressing IgE were assessed using anti-Mouse IgE PE (1:200, BioLegend, 406908; RME-1). Live/dead cells were discriminated after staining with Zombie Aqua viability dye. Cell proliferation was assessed using Cell Trace Violet according to manufacturer's instructions (CellTrace, Life Technologies).

Primary B cell reconstitution. Mature untouched B cells were purified as above and stimulated with LPS (5 μg/ml, Sigma, L7770) and mouse recombinant IL-4 (10 ng/ml, Peprotech, 214-14-20). Cultures were grown at 37 °C with 5% CO_2 under ambient oxygen conditions. Filtered retroviral supernatants, collected 48 h after co-transfection of BOSC23 cells with 7 μg pCL-Eco and 7 μg pMX-IRES-GFP-derived plasmids, were used to infect LPS/IL-4-stimulated B cell cultures in the presence of polybrene (2.5 μg/ml) and HEPES (20 mM) by spinoculation (850*g* for 90 min at 30 °C). After a rest period of 4 to 6 h, viral supernatants were removed, and replaced with LPS/IL-4-supplemented culture medium. Three days later, surface IgG1 expression was determined in populations of gated cells that were positive for the expression of an eGFP retroviral reporter.

CH12-F3 cell culture and CRISPR–Cas9 editing. All CH12-F3 cell lines were cultured in RPMI supplemented with 5% NCTC-109 medium, 10% FCS, 100 U/ml penicillin, 100 ng/ml streptomycin and 2 mM L-glutamine at 37 °C with 5% CO_2 under ambient oxygen conditions. *Rev7*$^{-/-}$, *c20orf196*$^{-/-}$ (*Shld1*$^{-/-}$), *Fam35a* (*Shld2*$^{-/-}$) and *Flj26957* (*Shld3*$^{-/-}$) CH12-F3 were generated using CRISPR–Cas9. In brief, gene-specific sgRNAs (sgRNA sequences in Extended Data Table 3a) were cloned in modified pX330 (Addgene #42230) or pX458 vectors (Addgene #48138). CH12-F3 cells were nucleofected (Amaxa Nucleofector 2b, Lonza) with 2 μg of plasmid and Cell Line Nucleofector Kit R (Lonza), using program D-023. Isogenic cell clones were isolated by limiting dilution (pX330) or GFP sorting (pX458) single cell into 96-well plates. Clones bearing bi-allelic indel mutations were identified by native PAGE resolution of PCR amplicons corresponding to edited loci (amplicon primer sequences in Extended Data Table 4b), and gene disruption subsequently confirmed by Sanger sequencing (sequencing results in Extended Data Table 3b). Where antibodies were available, effective target protein ablation was confirmed by immunoblotting. Complemented cell lines were generated by lentivirus-mediated transduction, using viral supernatants collected from 293T cells co-transfected with third generation packaging vectors and pLenti-PGK-PURO-DEST (Addgene #19068) or pLenti-PGK-Flag-HA-PURO-DEST vectors containing cloned transgene inserts. Typically, cells were spinoculated with polybrene (8 μg/ml) and HEPES (20 mM)-supplemented viral supernatants (1500 rpm, 90 min at 25 °C). Stable cell-lines were subsequently selected and maintained in the presence of puromycin (1 μg/ml). To stimulate CSR to IgA, CH12-F3 cells were stimulated with agonist anti-CD40 antibody (0.5 μg/ml; Miltenyi Biotec; FGK45.5), mouse IL-4 (5 ng/μl; R&D Systems) and TGFβ1 (2.5 ng/μl; R&D Systems). Cell-surface IgA expression was determined by flow cytometric staining with anti-mouse IgA-FITC antibody (Thermo Fisher; 11-4204-82; MA-6E1). CH12-F3 proliferation was monitored by dye dilution using carboxyfluorescein succinimidyl ester (CFSE) according to manufacturer's instructions (CellTrace; Life Technologies). In cell cycle experiments, CH12-F3 cells pulse-treated with BrdU for 30 min before fixation in 70% ethanol were stained with propidium iodide and rat anti-BrdU-FITC (1:100, Bio-Rad MCA2060FT). Certified mycoplasma free CH12-F3 were obtained by Cell Services (Francis Crick Institute). These and other cell lines (for example, 293T, BOSC-23) were confirmed free of mycoplasma contamination.

Antibodies. Immunoblot primary antibodies used in this study; anti-Rif1: Clone SK1316; gift of Ian Adams[30]; anti-histone H3: Abcam ab10799, clone 10799; 1:2,000; anti-Rev7: BD 612266, clone 14/MAD2B/Rev7, 1:500; anti-53BP1: Novus Biologicals NB100-304, 1:2,500; anti-HA-11: BioLegend 901501, Clone 16B12, 1:2,000; anti-actin: Sigma A1978, Clone AC15, 1:2,000; anti-tubulin: Sigma 00020911, Clone TAT-1, 1:10,000. Proteins were detected using HRP-conjugated secondary antibodies (α-rabbit, Thermo Fisher 31462, 1:50,000; anti-mouse, Thermo Fisher 31432, 1:50,000; Mouse TrueBlot ULTRA, Rockland 18-8817-33, 1:5000) and enhanced chemiluminescence (Clarity, Bio-Rad). Signals were acquired digitally on a Gel Doc XR system (Bio-Rad).

Proteomics and mass spectrometry. Pellets collected from cultures of ~4 × 10^7 CH12-F3 cells were lysed in BLB (Benzonase Lysis buffer: 20 mM HEPES pH 7.9, 40 mM KCl, 2 mM $MgCl_2$, 10% glycerol, 0.5% NP40, 50 U/ml Benzonase (Novagen), 0.05% (v/v) phosphatase inhibitors (Sigma-Aldrich) and protease

© 2018 Springer Nature Limited. All rights reserved.

inhibitors (Complete EDTA-free, Roche)) and were incubated on ice for 30 min before a second incubation with adjusted salt (450 mM KCl). Flag–REV7 or control complexes were isolated from clarified lysates, following their dilution in NSB (no-salt buffer: 20 mM HEPES (pH 7.9), 10% glycerol, 0.5 mM DTT, 0.5 mM EDTA, 0.05% (v/v) phosphatase inhibitors (Sigma-Aldrich) and protease inhibitors (Roche)) to a final salt concentration of 125 mM. Flag–HA–REV7 complexes were immunopurified on anti-Flag (M2) magnetic resin (Sigma-Aldrich), washed extensively in wash buffer (BLB supplemented with 125 mM KCl and 0.1% NP-40) and eluted with 3 × Flag peptide (Sigma-Aldrich). Flag-peptide eluted complexes were reduced and alkylated using DTT and iodoacetamide, followed by a chloroform/methanol precipitation. Proteins were resuspended in 6 M urea and digested over night with trypsin (Promega). Peptides were desalted (Sola, Thermo) and analysed on a LC–MS platform consisting of a Dionex Ultimate 3000 nHPLC and Q-Exactive mass spectrometer. Peptides were separated on an EASY-Spray column (50 cm, ES803, Thermo) with a gradient of 3–35% acetonitrile in 5% DMSO and 0.1% formic acid at 250 nl/min. MS1 spectra were acquired at a resolution of 70,000 at 200 *m*/*z*. Up to the 15 most abundant precursor ions were selected for subsequent MS/MS analysis after ion isolation with a mass window of 1.6 Th. Peptides were fragmented by HCD with 28% collision energy. Progenesis QI (v.3, Waters) was used for spectral counting and LFQ of the LC–MS/MS data with default parameters (top 3 quantitation mode). Proteins were identified with PEAKS 8.0 (Bioinformatics Solutions) using standard parameters and the Uniprot mouse reviewed proteome (retrieved 28 November 2017). Peptide false discovery rate (FDR) was set to 1% with a resulting protein FDR of 1.88%.

Chromatin immunoprecipitation. Each ChIP was performed from chromatin prepared from ~10^7 CH12-F3 cells stimulated for 30 h with agonist CD40 antibody, IL-4, and TGFβ1 as previously described[13]. Each ChIP was performed using 30–50 μg CH12-F3 chromatin using RPA34-20 (3 μg; Ab-3, Calbiochem) or anti-histone H3 (2 μg; ab1791, Abcam) coupled to 25 μl protein G Dynabeads (Life Technologies, 10003D). Relative quantities of ChIP-enriched DNA were calculated relative to total input chromatin by qPCR in triplicate on CFX96 Real-Time Analyzer (Bio-Rad) or StepOnePlus (Applied Biosystems) instruments using Quantifast SYBR Green reagent (Qiagen) and locus-specific primer pairs (Extended Data Table 4a).

Yeast two-hybrid interaction. Open reading frames encoding the indicated proteins were cloned into pAD-DEST and pBD-DEST vectors and transformed into *Saccharomyces cerevisiae* (strain PJ69-4A). Ten-millilitre cultures prepared from single transformants were diluted to equal volumes containing 2 × 10^7 cells, and fivefold serial dilutions 'spotted' on control (−Leu, −Trp) or experimental (−Leu, −Trp, −His) plates supplemented with 3-amino-1,2,4-triazole (3-AT, 6 mM, Sigma A8056). Plates were incubated at 30 °C for 3 days.

PARPi resistance. KB1P-G3 and *53bp1*$^{-/-}$ KB1P-G3 cells were infected with viral supernatant generated using lentiCRISPR-Bsr. Following selection in blasticidin (10 mg/ml), genomic DNA (gDNA) samples were collected immediately before seeding for olaparib sensitivity. Blasticidin-resistant populations were seeded in 6-well plates at a density of 10^4 cells per well (5 × 10^3 for *53bp1*$^{-/-}$ KB1P-G3) in the presence of olaparib or DMSO and grown at 37 °C (5% CO_2 and 3% O_2). Medium was refreshed at 4 days and 8 days. After 10 days, cultures were expanded in fresh medium for 1 week in 6-cm dishes before collecting gDNA. PCR amplicons encompassing the edited locus were PCR amplified from gDNA, Sanger sequenced (GATC Biotech) and analysed by tracking of indels by decomposition (https://tide.deskgen.com/). Surviving cells were collected and three replicates were plated in DMSO and three in olaparib for viability analysis. DMSO and olaparib-treated cells were stained with crystal violet (0.5% (w/v) crystal violet in 25% methanol) after 8 days and 10 days growth (37 °C, 5% CO_2 and 3% O_2), respectively. Crystal violet stained cells were dissolved in a 10% (v/v) acetic acid solution a minimum of 24 h after staining and the OD_{595} was measured as a quantitative metric of relative growth. In Fig. 4g, cells were incubated for 24 h in olaparib 250 nM in standard medium, before metaphase chromosomes were collected.

Cytological analysis. Metaphase spreads were prepared by standard methods. In brief, detached cells were resuspended in KCl 75 mM for 20 min, before being fixed in Carnoy's fixative. Approximately 20 μl of cell suspension was dropped onto clean slides and left to dry overnight. The cells were then stained with propidium iodide 0.5μg/ml in PBS for 20 min, rinsed and the slides mounted in Vectashield/DAPI (Vector Labs). The slides were analysed blind, and a minimum of 50 metaphases were acquired using an Olympus BX60 microscope for epifluorescence equipped with a Sensys CCD camera (Photometrics). Images were collected using Genus Cytovision software (Leica).

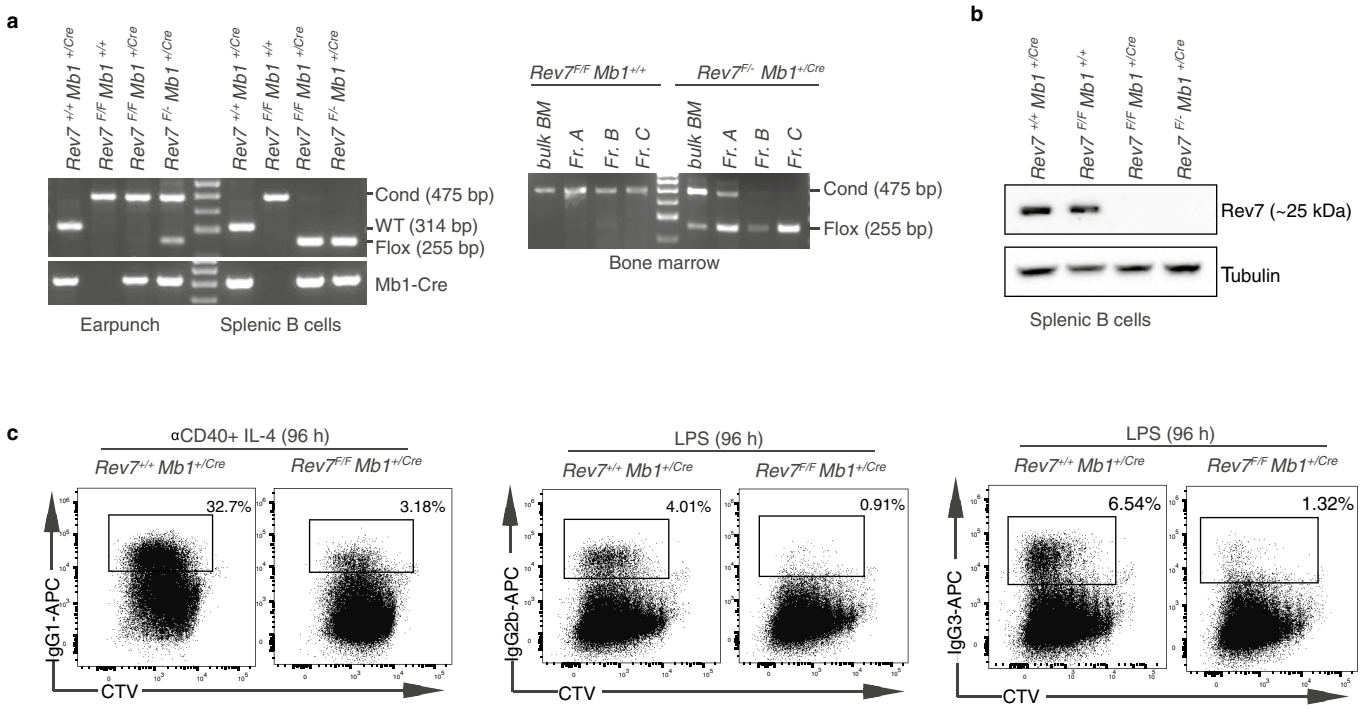
Fluorescence in situ hybridization analysis. BAC probes (RP23-41J14 (*Igh* 3′); RP24-316H6 (*Igh* 5′) (gifts from the Welcome Sanger Institute, Hinxton); RP23-374P12 (*Igl* 3′); RP23-382P9 (*Igl* 5′) (Source Bioscience)) were labelled using a nick translation kit (Abbott Molecular) according to the manufacturer's instructions, incorporating either Chromatide Alexa Fluor 594-5-dUTP (Thermo Fisher Scientific), Chromatide Alexa Fluor 488-5-dUTP (Thermo Fisher Scientific), Gold-dUTP (Abbott Molecular) or biotin-16-dUTP (Sigma). The probes were resuspended in the presence of a ten-times excess of unlabelled mouse C0t1 DNA (Thermo Fisher Scientific), in hybridization buffer (50% formamide, 10% dextran sulphate, 2×SSC), before being denatured for 8 min at 85 °C, followed by a pre-annealing step 30 min at 37 °C. The metaphase spreads were denatured in 0.07 N NaOH for 1 min. The probes were applied onto the slides, and the hybridization was carried out overnight at 37 °C. Three post-hybridization washes were performed, in 0.1× SSC buffer at 65 °C. Biotinylated probes were detected using streptavidin-Cy5 (Thermo Fisher Scientific). Slides were mounted in Vectashield/DAPI and analysed blind with the microscope described above. Between 98 and 150 metaphases were analysed for each mouse with the exception of one case (45 metaphases collected for one control mouse).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability statement. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository, with the dataset identifier PXD009650 (https://doi.org/10.6019/PXD009650).

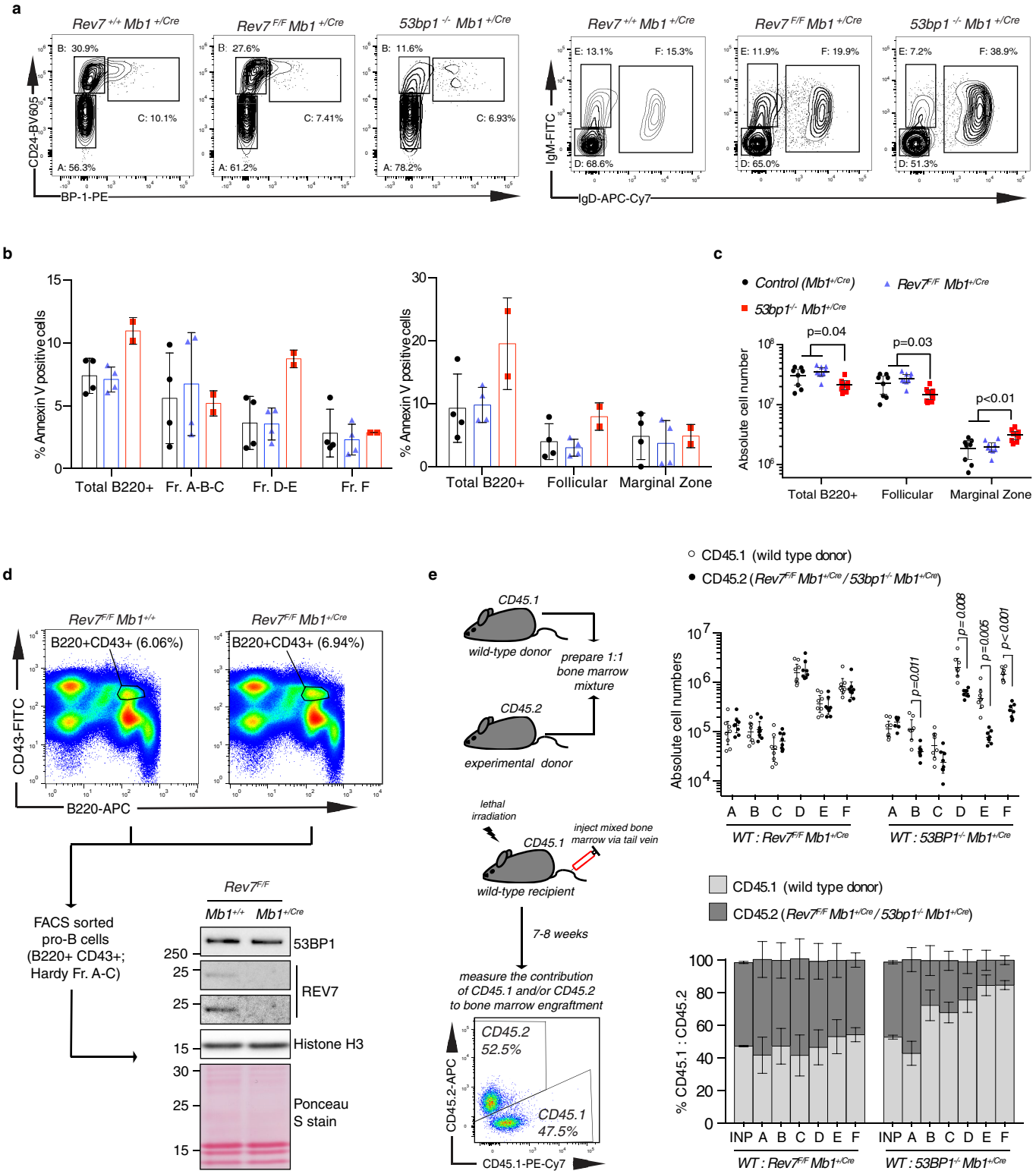
29. Ward, I. M., Minn, K., van Deursen, J. & Chen, J. p53 binding protein 53BP1 is required for DNA damage responses and tumor suppression in mice. *Mol. Cell. Biol.* **23**, 2556–2563 (2003).
30. Adams, I. R. & McLaren, A. Identification and characterisation of mRif1: a mouse telomere-associated protein highly expressed in germ cells and embryo-derived pluripotent stem cells. *Dev. Dyn.* **229**, 733–744 (2004).
31. Skarnes, W. C. et al. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337–342 (2011).
32. Khalaj, M. et al. A missense mutation in Rev7 disrupts formation of Polζ, impairing mouse development and repair of genotoxic agent-induced DNA lesions. *J. Biol. Chem.* **289**, 3811–3824 (2014).
33. Kikuchi, S., Hara, K., Shimizu, T., Sato, M. & Hashimoto, H. Structural basis of recruitment of DNA polymerase ζ by interaction between REV1 and REV7 proteins. *J. Biol. Chem.* **287**, 33847–33852 (2012).
34. Wagner, S. A. et al. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol. Cell. Proteomics* **10**, M111.013284 (2011).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | CSR characterization in *Rev7* conditional-knockout mice. a, PCR amplicons from genomic DNA obtained by ear biopsies or purified splenic B cells (left), or flow-cytometry-sorted cells from Hardy fractions A, B and C (right) from mice of the indicated genotype. Bands of different size correspond to the $Rev7^{tm1c}$ allele (cond, 475 bp), $Rev7^{+}$ allele (wild type, 314 bp) and $Rev7^{tm1d}$ allele (flox, 255 bp). Representative data; $n > 3$ experiments. **b**, Western blot analysis of REV7 protein expression in splenic B cells isolated from mice with the indicated genotype. Representative data; $n = 2$ experiments. For gel source data, see Supplementary Fig. 1. **c**, CTV-labelled purified B cells were stimulated as indicated and stained for surface IgG1 (left) or IgG2b (centre) or IgG3 (right) on day 4. Representative of $n > 6$ experiments.

© 2018 Springer Nature Limited. All rights reserved.

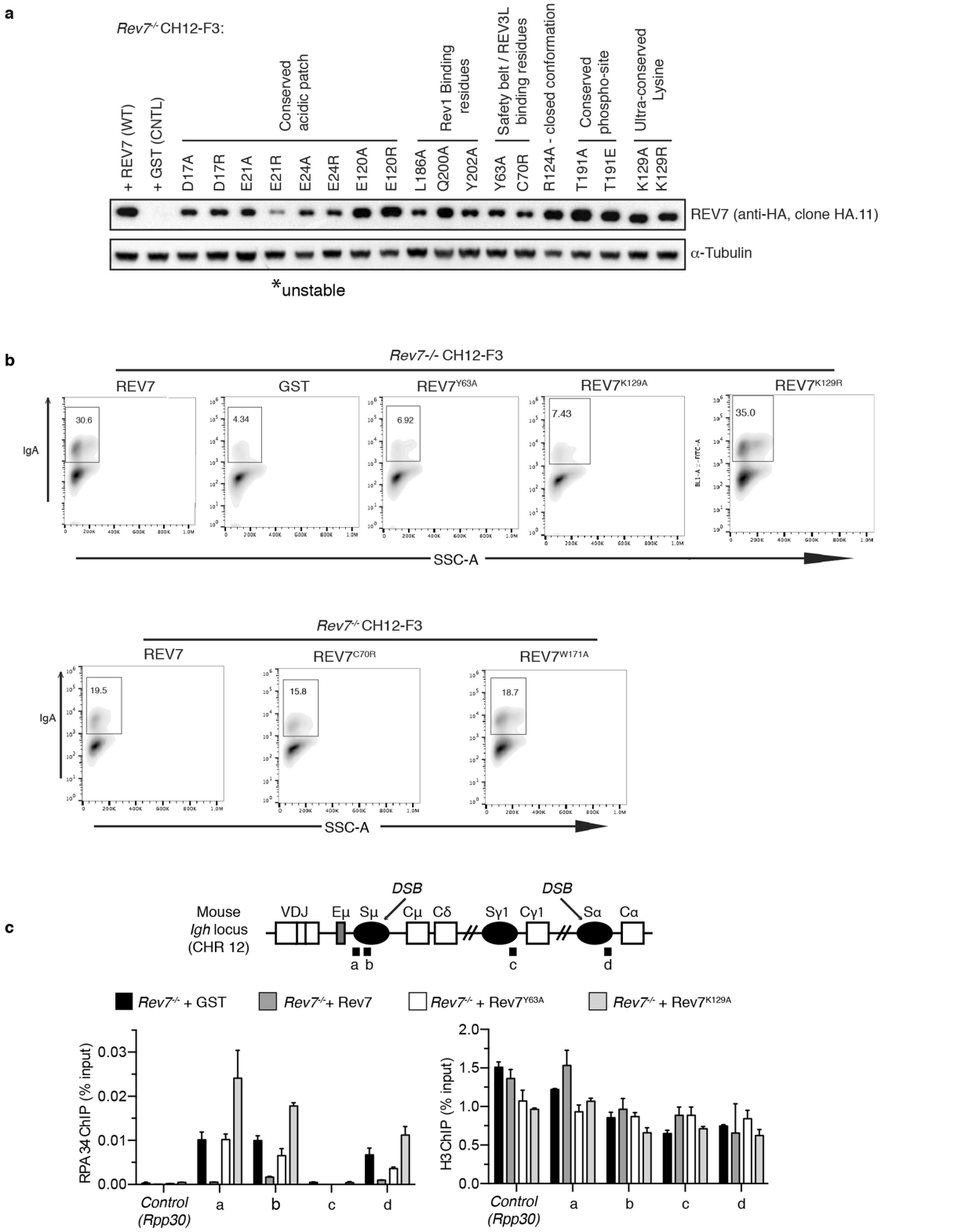


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

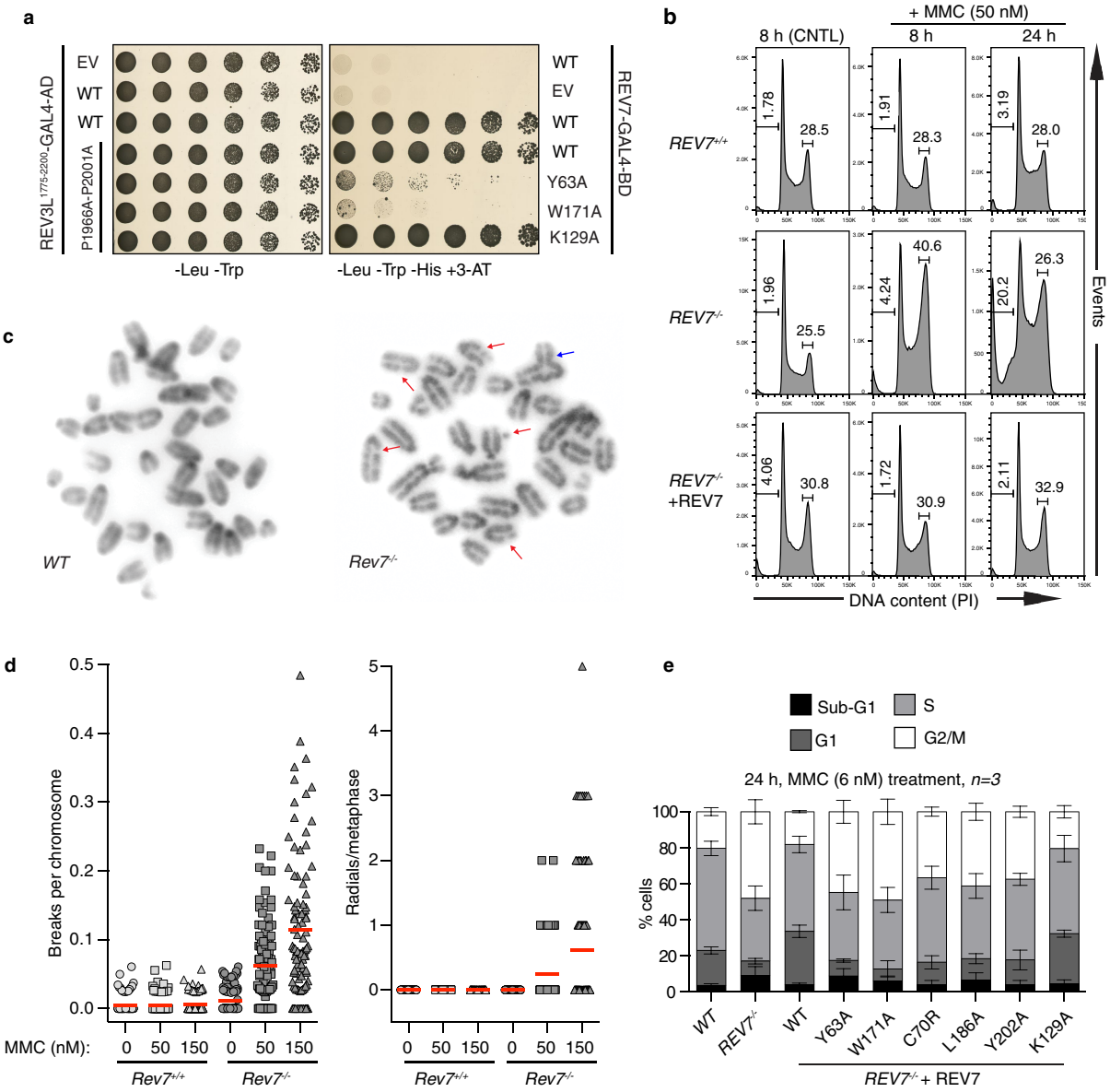
Extended Data Fig. 2 | B cell lineage developmental differences in *Rev7*- and *53bp1*-deficient mice. a, Flow cytometry analysis of B cell development in the bone marrow of $Rev7^{+/+}Mb1^{+/cre}$, $Rev7^{f/f}Mb1^{+/cre}$ and $53bp1^{-/-}Mb1^{+/cre}$ mice; gating on $B220^{+}CD43^{+}$ (left, Hardy fractions A, B and C) and on $B220^{+}CD43^{-}$ (right, Hardy fractions D, E and F). Representative data; $n > 8$ experiments. **b**, Apoptotic indices of total, pre-pro-B to pre-B, immature and mature B cell fractions in the bone marrow (left) and spleen (right) of $Rev7^{+/+}Mb1^{+/cre}$ ($n = 4$), $Rev7^{f/f}Mb1^{+/cre}$ ($n = 4$), $53bp1^{-/-}Mb1^{+/cre}$ ($n = 2$) mice. **c**, Flow cytometric sub-classification of mature splenic B cell fractions in $Rev7^{+/+}Mb1^{+/cre}$ ($n = 8$), $Rev7^{f/f}Mb1^{+/cre}$ ($n = 8$), $53bp1^{-/-}Mb1^{+/cre}$ ($n = 9$) mice. *P* values, unpaired Student's two-tailed *t*-test. Bars represent mean ± 95% confidence interval. **d**, Top, indicated populations of pooled pro- to pre-B cell stage bone marrow lymphocytes ($B220^{+}CD43^{+}$; Hardy fraction A, B and C) from $n = 2$ mice per genotype were FACS-sorted and used to generate whole cell extracts. Bottom, immunoblot shows an absence of REV7 protein in extracts prepared from $Rev7^{f/f}Mb1^{+/cre}$ experimental bone marrow, when compared to extracts prepared from $Rev7^{f/f}Mb1^{+/+}$ (no Cre) controls, yet equivalent levels of 53BP1, histone H3 (loading control) and total protein (Ponceau S stain). For gel source data, see Supplementary Fig. 1. **e**, Left, diagram of the mixed bone marrow chimaera transplantation experiment. Bone marrow cells from a wild-type $CD45.1^{+}$ donor mouse were combined with an equal number of bone marrow cells from an experimental $CD45.2^{+}$ donor mouse, and injected into lethally irradiated recipient $CD45.1^{+}$ mice ($n = 8$ per genotype). After eight weeks, the recipient bone marrow was analysed for the relative contribution of $CD45.1^{+}$ or $CD45.2^{+}$ cells to reconstitute the recipient mice. Right top, enumeration of B cell precursors (as per Fig. 1g) of $CD45.1^{+}$ (white circles) or $CD45.2^{+}$ (from $Rev7^{f/f}Mb1^{+/cre}$ or $53bp1^{-/-}Mb1^{+/cre}$ mice; black circles) cells in the bone marrow after reconstitution. Right bottom, stage-specific ratios of $CD45.1^{+}$ to $CD45.2^{+}$ grafted B cells for indicated mixed chimaeras. In parallel, an additional control experiment involving wild-type CD45.1 and $Mb1^{+/cre}$ mixed chimaera was performed, resulting in equal CD45.1:CD45.2 reconstitution (data not shown). *P* values, multiple *t*-test with Holm–Sidak correction; mean ± 95% confidence interval. INP, input.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | REV7 protein expression and CSR in complemented *Rev7*$^{-/-}$ CH12-F3 cells. a, Stable Flag–HA–REV7 protein expression in indicated complemented CH12-F3 cell lines as determined by western blotting with HA-specific antibody. Tubulin served as loading control. Representative data; $n > 3$ independent experiments. For gel source data, see Supplementary Fig. 1. **b**, Flow-cytometric plots of CSR data in Fig. 2b. Representative data; $n > 3$ independent experiments. **c**, Chromatin prepared from indicated CH12-F3 lines 30 h after stimulation was subjected to ChIP with RPA34 or histone H3 (control) antibodies. Schematic of mouse *Igh* locus with positions of qPCR amplicons. ChIP recoveries were quantified against input DNA. Representative data, $n = 2$ independent experiments. Mean ± s.d.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Distinct REV7 residues mediate its functions in interstrand cross-link repair. a, Indicated REV7 mutants were tested for their interaction with the RBM_1 of REV3L by yeast two-hybrid assay. A fragment of REV3L (amino acids 1775–2200) with RBM_2 mutated (P1996A and P2001A) was used. Fivefold dilutions, $n = 4$ biological repeats. **b**, Representative histograms of indicated CH12-F3 lines, mock- or MMC-treated for 8 and 24 h. Numbers depict sub-G1 and G2/M populations as a proportion (%) of total events. Representative data, $n = 3$ independent experiments. **c**, Representative propidium iodide and DAPI co-stained metaphases from indicated CH12-F3 cell lines after 24 h incubation with MMC; $n = 2$ independent experiments. Arrows, chromosome breaks (red) and radial chromosome (blue). **d**, Blind quantification of DSBs (chromatid and chromosome-type) and radials in indicated cell lines, with each metaphase scored as a single point. $n = 2$ independent experiments, each with 50 metaphases scored per genotype and condition. Bars indicate mean. **e**, Cell cycle of indicated CH12-F3 lines after MMC treatment. Proportions (%) of sub-G1, G1, S and G2/M events. $n = 3$ independent experiments ± s.d.

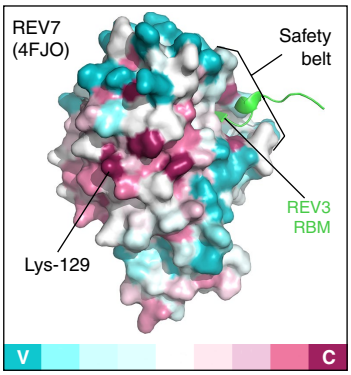
© 2018 Springer Nature Limited. All rights reserved.

a

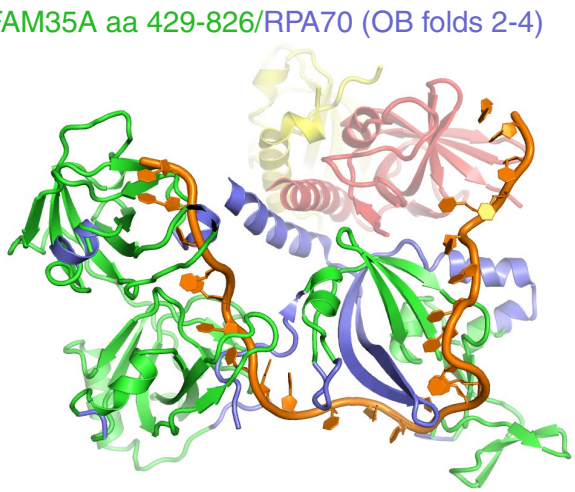
```
HUMAN REV3L aa1876-1889 NILKPLMSPPSREE
MOUSE REV3L aa1873-1886 HILKPLMSPPSREE
CHICK REV3L aa1780-1793 HILKPLMSPPSREE
DANRE REV3L aa1538-1551 RILKPLMSPPAREE
HUMAN FLJ26957 aa7-20    LHYRPCESDPTQLP
BOVIN FLJ26957 aa7-20    LHYRPYESDPTQLP
MOUSE FLJ26957 aa12-25   LHYRPYENDPKQLA
CHICKEN FLJ2695 aa5-18   LHYRPHQRDLTELQ
CONSENSUS:                [K/R]PxxxxP

HUMAN REV3L aa1876-1889 NILKPLMSPPSREE
MOUSE REV3L aa1873-1886 HILKPLMSPPSREE
CHICK REV3L aa1780-1793 HILKPLMSPPSREE
DANRE REV3L aa1538-1551 RILKPLMSPPAREE
HUMAN FLJ26957 aa48-62   LPLRPKRSPPVISE
BOVIN FLJ26957 aa48-62   LPLKPKRSPPGISE
MOUSE FLJ26957 aa53-67   LPLKPERLPPVISE
CHICK FLJ26957 aa46-60   LPLKPKKQPPVISS
CONSENSUS:              L[K/R]PxxxPP
```

b



c



d

```
FAM35A iso2  429 TSLIKNCDSKSQKYNCLVMVLSPCHVREINIKFGPNSGSKVPLATVTVIDQSETKKKVFLIRTAA-FVAFTVFLGDIILL
umRPA70      182 IYPIEGLSPYQNRWTIKARVTSKSDIRHWSNQRGE-----GKLFSVNLLDDSG-EIKATGFNDAVDRFYPLLQENHVYLI

FAM35A iso2  508 TDVVIHEDQ-----WIGETVLQSTFSSQLLNLGSYSSIQPEEYSSVVSEVVLQDLLAYVSSKHSYLRDLPPRQPQFVNSI
umRPA70      256 SKARVNIAKKQFSNLQNEYEITFENSTEI----------------------------------EECTDATDVPEVRYEFV

FAM35A iso2  583 DFVELEHLQPDVLVHAVLRVVDFTILTEAVYSYRGQKQKKVMLTVEQAQDQHYALVLWGPGAAWYPQLQR-KKGYIWEFK
umRPA70      302 RINELESVEANQQCDVIGILDSYGELSEIVSKASQRPVQKRELTLVDQGNRSVKLTLWGKTAETFPTNAGVDEKPVLAFK

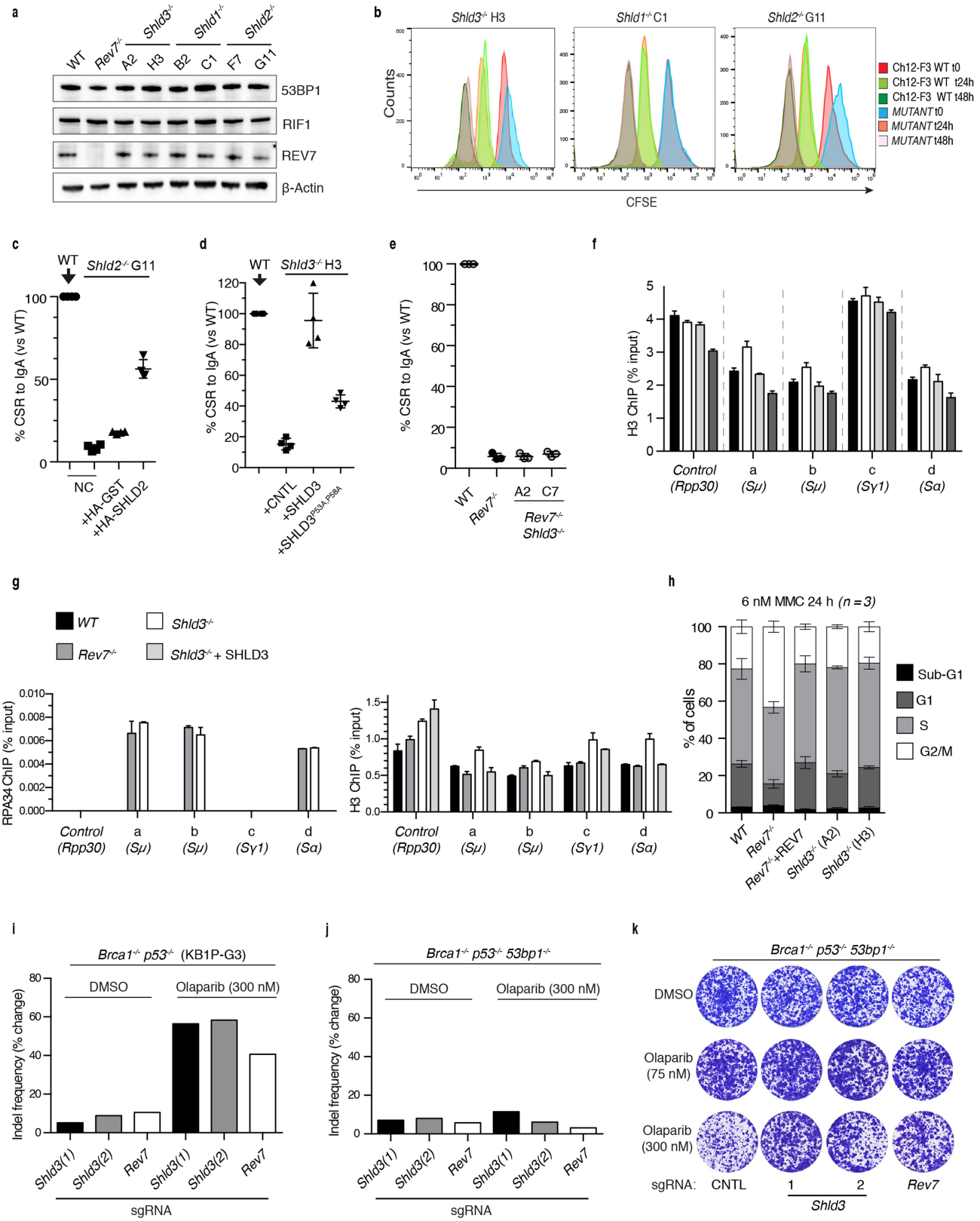
FAM35A iso2  662 YLFVQCNYTLENLEIHTTPWSSCECTFDDDIRAITFKAKFQKSAPSFVKI---------------SDFATHLEDKCS---
umRPA70      382 GVKVGDF---GGRSLSMFSSSTML-LNPDITESHVLRGWYDNDGAHAQFQPYTNGGGAGANMAERRTLVQVKDENLGMSE

FAM35A iso2  724 --GVVLLKAQISEEAFPITASQKIALNAHSSLKSIFSSLPNLVYTGCAK--CGLELETDENRIYKQCFSCLPFTMKKIYY
umRPA70      466 KPDYFNVRATVVYLKQE-----------------------NLYYTACASEGCNKKVNLDHENNWR-CEKCDRSYATPEYR

FAM35A iso2  800 RPALMTAIDGRHDVCIRVESKLIEKILLNISADCLNR 836
umRPA70      523 YILSTNVADATGQLWLSGFNEDATQL--IGMSAGELHK 558
```

Extended Data Fig. 5 | Identification of the REV7–shieldin complex. **a**, N-terminal SHLD3 (FLJ26957) sequences contain conserved REV7-binding motifs (RBM_1 and RBM_2). Pairwise alignment of RBM_1 of REV3L with RBM_1 and RBM_2 of SHLD3 across multiple species. Black and grey shading indicate identical and similar residues, respectively. **b**, Surface representation of REV7 crystal structure (RCSB Protein Data Bank (PDB) code: 4FJO), pseudo-coloured according to amino acid conservation. V, variable residues; C, conserved residues. **c**, Protein threading model of FAM35A identifies an RPA70-like triple OB-fold architecture. The FAM35A structural model generated using Phyre2 predicts that amino acid residues 429–826 (coloured in green) adopt the same folds as *Ustilago maydis* RPA70[24] (PDB code: 4GOP, chain C; 99.5% confidence in the model[23]). Blue structural chains from RPA70 are inserted in regions of FAM35A that could not be modelled with high confidence (internal regions and FAM35A residues above 826). In the shieldin complex, RPA34 (opaque red) and RPA14 (opaque yellow) subunits of the heterotrimeric RPA complex might be substituted with SHLD1, SHLD3 and REV7 proteins. **d**, Pairwise alignment shows 16% sequence identity across human FAM35A isoform 2 (OB-folds 1 to 3) and *U. maydis* RPA70 (OB-folds 2 to 4; generated with Phyre2). Black and grey shading indicate identical and similar residues, respectively.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 6 | SHLD3 (FLJ26957) mediates 53BP1-dependent NHEJ. a, Immunoblot showing levels of REV7, 53BP1 and RIF1 protein in whole cell lysates prepared from indicated CH12-F3 cell lines. Representative of $n = 2$ individual experiments. For gel source data, see Supplementary Fig. 1. **b**, Normal proliferation of stimulated wild type and *Shld3*$^{-/-}$, *Shld1*$^{-/-}$ and *Shld2*$^{-/-}$ CH12-F3 cells. CFSE dye dilution assay. $n = 2$ independent experiments. **c**, Summary of IgM-to-IgA CSR frequencies in indicated control GST or HA–SHLD2(mouse)-complemented *Shld2*$^{-/-}$ CH12-F3 cells. Data normalized to CSR in wild-type cells. $n = 4$ independent experiments; mean ± s.d. NC, non-complemented. **d**, IgM-to-IgA CSR in indicated *Shld3*$^{-/-}$CH12-F3 cells ±complementation with wild-type SHLD3 or SHLD3(P53A, P58A) (that is, RBM_2 mutated). Data normalized to CSR in wild-type cells. Mean ± s.d. ($n = 4$). **e**, Summary of IgM-to-IgA CSR frequencies in *Rev7*$^{-/-}$ and *Rev7*$^{-/-}$*Shld3*$^{-/-}$ double knockout CH12-F3 clones. Data normalized to CSR in wild-type cells. $n = 4$ independent experiments; mean ± s.d. **f**, Histone H3 (control) ChIP efficiencies at indicated *Igh* and control loci. This panel is related to RPA ChIP data shown in Fig. 4d. Representative data, $n = 2$ independent experiments, mean ± s.d., 2 or 3 qPCR replicates. **g**, Indicated CH12-F3 lines stimulated with anti-CD40 antibody, IL-4 and TGFβ1 (30 h) were subjected to ChIP experiments with RPA34 and histone H3 (control) antibodies. Representative data, $n = 2$ independent experiments, mean ± s.d., 2 or 3 qPCR replicates. **h**, Cell cycle of indicated CH12-F3 lines after MMC treatment. Proportions (%) of sub-G1, G1, S and G2/M events. $n = 3$ independent experiments; mean ± s.d. **i**, Change (%) of Cas9-dependent indels at the indicated sgRNA locus in KB1P-G3 cells after outgrowth in DMSO or olaparib (300 nM) for 7 days. Representative data, 2 independent experiments. **j**, As in **i** but with *53bp1*$^{-/-}$ KB1P-G3 cells. Representative data, $n = 2$ independent experiments. **k**, As in Fig. 4e but with *53bp1*$^{-/-}$ KB1P-G3 cells. Representative data, $n = 3$ independent experiments.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Predicted and observed offspring of *Rev7* mutant mice (*Rev7*$^{tm1a(EUCOMM)Wtsi}$; Mouse Genome Informatics code:4432091) bred in this study on a C57Bl/6 background

Cross:		*Rev7*$^{+/-}$ x *Rev7*$^{+/-}$	
Offspring (IMPC[31])		136	
Offspring (This study)		25	
Combined offspring		161	
Genotypes	*Rev7*$^{+/+}$	*Rev7*$^{+/-}$	*Rev7*$^{-/-}$
Predicted offspring	40.25	80.5	40.25
Observed offspring	52	109	0
Predicted offspring (corrected for homozygous lethality):	53.67	106.26	0

IMPC, International Mouse Phenotyping Consortium[31].

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Mutant mouse REV7 proteins generated and phenotyped in this study

Name	REV7 mutants	Description	References
Safety belt	$REV7^{Y63A}$ $REV7^{W171A}$ $REV7^{Y63A,W171A}$	Predicted to block REV7-REV3L interactions and sensitizes DT40 cells to Cisplatin	17
Mouse mutant alleles	$REV7^{C70R}$	Mouse missense mutation predicted to disrupt DNA polymerase ζ formation; sensitizes cells to ICLs.	32
Rev1-binding residues	$REV7^{L186A}$ $REV7^{Q200A}$ $REV7^{Y202A}$ $REV7^{L186A,Q200A}$	Predicted to block REV7-REV1 interactions	16,33
PTM/ Degron	$REV7^{R6A,L9A}$	D-box mutation blocks APC/C-dependent Rev7 degradation	18
	$REV7^{K44R, K90R, K162R, K167R, K190R}$	Identified REV7 ubiquitinated sites	34 Phosphosite.org
	$REV7^{K44R, K46R, K47R, K72R, K77R, K82R, K90R, K97R, K129R, K162R, K167R, K190R, K198R, K208R}$	Lysine-less REV7 mutant	
	$REV7^{T191A}$ $REV7^{T191E}$	Reported phosphorylation site.	Phosphosite.org
Conserved surface residues	$REV7^{K129A}$ $REV7^{K129R}$	Conserved Lysine	Identified by structure-led conservation analysis
	$REV7^{D17A}$ $REV7^{D17R}$ $REV7^{E21A}$ $REV7^{E21R}$ $REV7^{E24A}$ $REV7^{E24R}$ $REV7^{E120A}$ $REV7^{E120R}$	Conserved acidic patch	

Further information about the REV7 mutants can be found in refs.[16–18,32–34].

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 3 | CRISPR–Cas9 reagents and edited cell-lines generated in this study

a

Target gene name / sgRNA #		Sequence (5'-3')	Description
Rev7 (Mad2l2/Mad2b)		CCTGATTCTCTATGTGCGCG	Targeted to exon 4, S
C20orf196 / Shld1	1	CTGTACCTTGGATCTACCCG	Targeted to exon 2, S
	2	TCTTCTTCGACTACTGATCC	Targeted to exon 2, S
Fam35a / Shld2	1	TCGGACTTTTTCCCAGGCGT	Targeted to exon 4, AS
	2	CAGGGAGTCAGACGAACTGA	Targeted to exon 4, AS
	3	ACTGCATGGACTAATATATC	Targeted to exon 6, AS
Flj26957 / Shld3	1	GACTCATCGTATGGAAACCA	Targeted to exon 1, AS
	2	GGAAGTTTGGACTCATCGTA	Targeted to exon 1, AS

b

	Target gene	Clone	Genotype	
			Allele 1	Allele 2
CH12-F3	***Rev7***	C	+1 bp at R34	-2 bp at R34
	C20orf196 / Shld1	B2	+2 bp at L20	+2 bp at L20
		C1	-14 bp at S54	+1 bp at D58
		D1	-5 bp at T57	-5 bp at T57
	Fam35a / Shld3	E4	-1 bp at A35	+32 bp at A35
		F7	-25 bp at T32	-19 bp at N34
		G11	-5 bp at R273	-20 bp at G270
	Flj26957 / Shld3	A2	-1 bp at W46	-1 bp at W46
		A4	- 2 bp at F42	-7 bp at I44
		F9	-2 bp at Y49	-2 bp at Y49
		H3	-4 bp at Y49	-13 bp at K53
	Flj26957 in Rev7$^{-/-}$ (C)	A2	-23 bp at F38	-23 bp at F38
		C7	-22 bp at F38	-22 bp at F38
KB1P-G3 *(Brca1$^{-/-}$ p53$^{-/-}$)*	***Flj26957***	7	-1 bp at W46	-8 bp at W46
		8	-1 bp at W46	-5 bp at W46

a, Sequences of sgRNA used in gene editing experiments. S indicates sgRNA corresponds to sense strand; AS indicates antisense strand. **b**, Individual edited alleles in each cell-line clone as confirmed by Sanger sequencing.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 4 | Primers used in this study

a

Target/gene locus		Primer Sequence (5'-3')
Rpp30	Fwd	TCCAGTGTGCAAGAAAGCTAAATG
	Rev	GGCAGTGCGTGGAGACTCA
A (*Igh Sμ*)	Fwd	CAATGTGGTTTAATGAATTTGAAGTTGCCA
	Rev	TCTCACACTCACCTTGGATCTAAGCACTGT
B (*Igh Sμ*)	Fwd	GCTAAACTGAGGTGATTACTCTGAGGTAAG
	Rev	GTTTAGCTTAGCGGCCCAGCTCATTCCAGT
C (*Igh Sγ1*)	Fwd	AGTGTGGGAACCCAGTCAAA
	Rev	GTACTCTCACCGGGATCAGC
D (*Igh Sα*)	Fwd	TGAAAAGACTTTGGATGAAATGTGAACCAA
	Rev	GATACTAGGTTGCATGGCTCCATTCACACA

b

Target gene	gRNA	Sequence (5'-3')	
C20orf196 / Shld1	1	Fwd	AGTAGCTGCTCTTTTGGCGT
		Rev	TGGCAGCAGAAAGAGAAGGG
	2	Fwd	AGTAGCTGCTCTTTTGGCGT
		Rev	TGGCAGCAGAAAGAGAAGGG
Fam35a / Shld2	1	Fwd	AGCCCACACATTTGTCCACT
		Rev	GCTCCAGTTGCTCCACTGAA
	2	Fwd	TCCTTGGCTTCTTGGACACC
		Rev	ATGGAGGCAGAACCAACAGG
	3	Fwd	TGACCTTGAGCCTGTTCCAC
		Rev	TGGAGTTGGAGCAGTTGCAG
Flj26957 / Shld3	1	Fwd	TACTGCTTCACGCTCTCAGC
		Rev	GCAGGCTGTCCCTACCAAAT
	2	Fwd	TACTGCTTCACGCTCTCAGC
		Rev	GCAGGCTGTCCCTACCAAAT

a, Locus-specific ChIP amplicon qPCR primer sequences. **b**, Primers used to amplify CRISPR–Cas9 target loci in edited cell lines.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0308-7

Structures of human Patched and its complex with native palmitoylated sonic hedgehog

Xiaofeng Qi[1,5], Philip Schmiege[1,5], Elias Coutavas[2], Jiawei Wang[3]* & Xiaochun Li[1,4]*

Hedgehog (HH) signalling governs embryogenesis and adult tissue homeostasis in mammals and other multicellular organisms[1–3]. Whereas deficient HH signalling leads to birth defects, unrestrained HH signalling is implicated in human cancers[2,4–6]. N-terminally palmitoylated HH releases the repression of Patched to the oncoprotein smoothened (SMO); however, the mechanism by which HH recognizes Patched is unclear. Here we report cryo-electron microscopy structures of human patched 1 (PTCH1) alone and in complex with the N-terminal domain of 'native' sonic hedgehog (native SHH-N has both a C-terminal cholesterol and an N-terminal fatty-acid modification), at resolutions of 3.5 Å and 3.8 Å, respectively. The structure of PTCH1 has internal two-fold pseudosymmetry in the transmembrane core, which features a sterol-sensing domain and two homologous extracellular domains, resembling the architecture of Niemann–Pick C1 (NPC1) protein[7]. The palmitoylated N terminus of SHH-N inserts into a cavity between the extracellular domains of PTCH1 and dominates the PTCH1–SHH-N interface, which is distinct from that reported for SHH-N co-receptors[8]. Our biochemical assays show that SHH-N may use another interface, one that is required for its co-receptor binding, to recruit PTCH1 in the absence of a covalently attached palmitate. Our work provides atomic insights into the recognition of the N-terminal domain of HH (HH-N) by PTCH1, offers a structural basis for cooperative binding of HH-N to various receptors and serves as a molecular framework for HH signalling and its malfunction in disease.

The HH precursor undergoes autocatalytic processing in the endoplasmic reticulum to release an amino-terminal signalling domain (HH-N) with cholesterol covalently coupled to its carboxyl terminus. Hedgehog acyltransferase then adds palmitate to the α-amino group of the N-terminal-specific cysteine to yield the mature, doubly lipidated signalling molecule[9,10]. N-terminal palmitoylation is indispensable for HH signalling: (1) fatty-acylated SHH-N is far more active than unacylated SHH-N, as determined from differentiation assays and HH signalling assays[11,12]; (2) blocking HH-N palmitoylation (by mutation of its palmitoylation site) was shown to affect embryonic development in *Drosophila* and in mice[10,13]; and (3) inhibitors of hedgehog acyltransferase that prevent the palmitoylation of SHH block HH signalling[14].

Human patched 1 (PTCH1), the primary receptor for HH-N ligands, consists of 1,447 amino acids, including 12 transmembrane helices and three approximately 30-kD soluble domains, namely two extracellular domains (ECD-I and ECD-II) that bind HH-N and one cytoplasmic carboxyl-terminal domain (CTD) (Fig. 1a, Extended Data Fig. 1). In addition, transmembrane helices 2–6 (TM2–TM6) of Patched are predicted to form a sterol-sensing domain (SSD), which, in other proteins such as NPC1 and HMG-CoA reductase, is involved in cholesterol metabolism and signalling[15]. Unliganded Patched inhibits HH signalling and this repression is released when HH binds to Patched[16]. Specifically, after HH binding, Patched releases its inhibition of SMO, a polytopic membrane receptor that activates the Gli transcription factors to upregulate HH target genes[2]. How Patched inhibits SMO is unknown, but there are studies that show that Patched may act indirectly by releasing a small molecule to regulate SMO[17,18]. In support of this model, Patched has a similar transmembrane topology to NPC1 and prokaryotic resistance–nodulation–cell division (RND) transporters, which transport ligands across membranes[7,19].

Two major gaps remain in our knowledge of the HH pathway: (1) the molecular details of how HH recognizes and binds Patched; and (2) the mechanism of SMO activation after HH binds Patched. More importantly, Patched is a tumour suppressor involved in basal cell carcinoma,

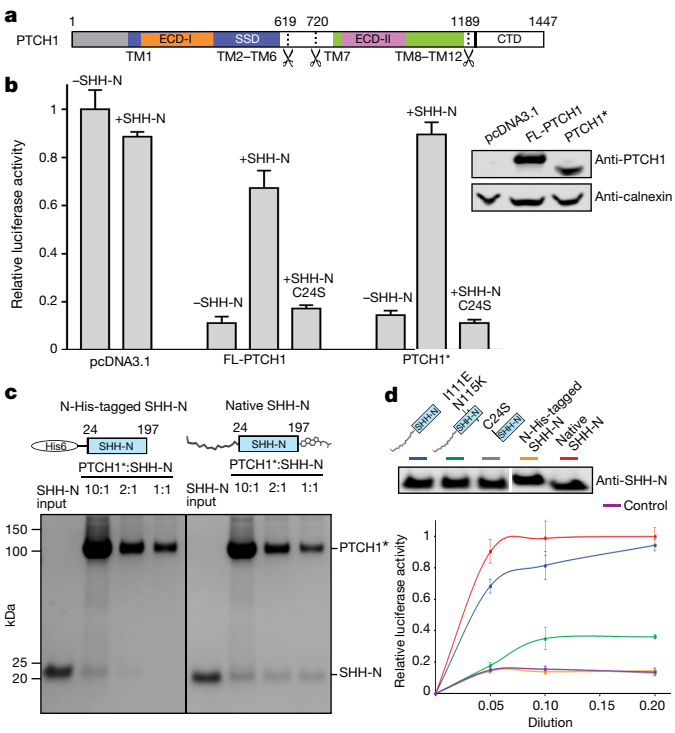


Fig. 1 | The engineered human PTCH1 protein binds a SHH-N ligand. **a**, Primary structure of PTCH1. Residues 619–720 and 1189–1447 were removed in PTCH1*. **b**, HH signalling in *Ptch1*$^{-/-}$ mouse embryonic fibroblasts (MEFs) transfected with pcDNA3.1, full-length PTCH1 (FL-PTCH1) or PTCH1* and response to wild-type or C24S mutant SHH-N ligand via luciferase activity. SHH-N in conditioned medium and transiently expressed PTCH1 were detected by western blotting. Calnexin served as an internal control and was detected by anti-calnexin antibody. SHH-N or SHH-N C24S was added to stimulate HH signalling. **c**, Pull-down assay of N-His-tagged SHH-N (left) or native SHH-N (right) with PTCH1* at different molar ratios detected by Coomassie staining. The assay was reproduced three times with similar results. **d**, Palmitoylated SHH-N stimulates HH signalling, but unmodified SHH-N does not. SHH Light II cells were treated with various concentrations of SHH-N variants, and HH signalling was measured using luciferase activity. Data (**b** and **d**) are mean ± s.d. ($n = 3$).

[1]Department of Molecular Genetics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Laboratory of Cell Biology, The Rockefeller University, New York, NY, USA. [3]State Key Laboratory of Membrane Biology, School of Life Sciences, Tsinghua University, Beijing, China. [4]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [5]These authors contributed equally: Xiaofeng Qi, Philip Schmiege. *e-mail: jwwang@tsinghua.edu.cn; xiaochun.li@utsouthwestern.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

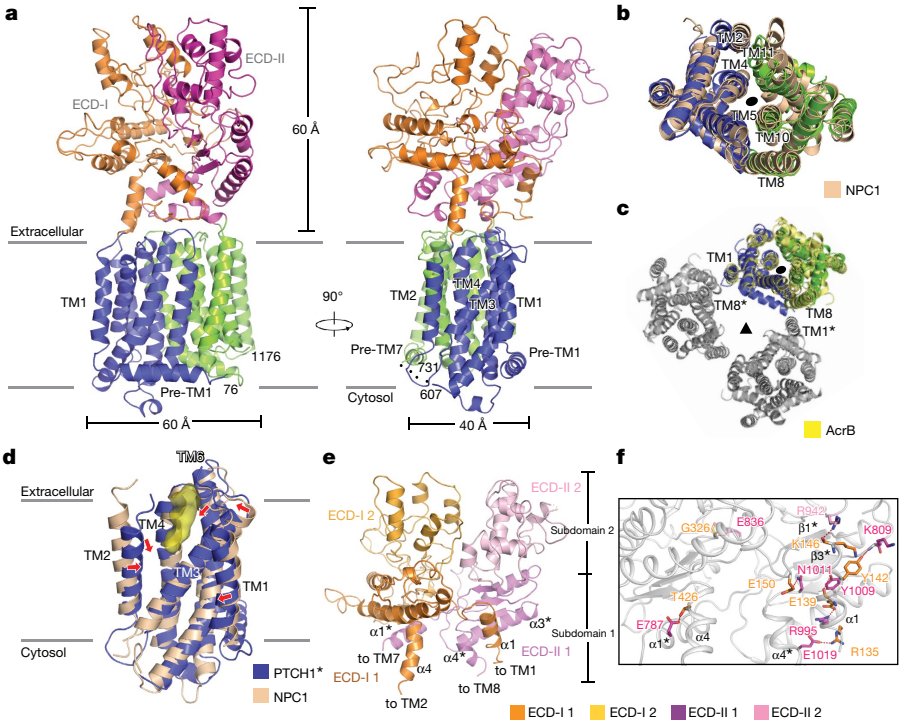


Fig. 2 | Overall structure of PTCH1*. a, Ribbon representation of the structure horizontal to the membrane. Flexible linkers are indicated by dots. **b**, Structural comparison of transmembrane domains of PTCH1* (blue and green) and NPC1 (brown; PDB code 5U74) viewed from the extracellular side. **c**, Structural comparison of transmembrane domains of PTCH1* (blue and green) and an AcrB trimer (PDB code 1IWG). One subunit of AcrB is yellow; the rest are grey. **d**, SSD comparison of PTCH1* (blue) and NPC1 (brown; PDB code 5U74) in a similar view to the right panel of **a**. Red arrows indicate shifted helices. A surface representation of the unidentified molecule is shown in yellow. **e**, Overall structure of ECD-I and ECD-II. **f**, Interface between ECD-I and ECD-II. Hydrophilic interactions are indicated by dots and residues are coloured as in **e**.

medulloblastoma and primitive neuroectodermal tumours[5], and mutations of SHH and PTCH1 can also cause developmental defects[3]. SMO is a target of antitumour agents[20]. Because this pathway is associated with human diseases, structural knowledge of Patched and the Patched–HH complex is crucial not only for elucidating the mechanism of signal transduction, but also for understanding the pathology of mutants and for the development of potential therapeutics for human diseases.

The full-length human PTCH1, expressed with a C-terminal Flag tag in human embryonic kidney (HEK)-293S cells, is eluted in the void volume during gel filtration (Extended Data Fig. 2a). To make the protein amenable to structural studies, we truncated the cytoplasmic loop between TM6 and TM7 and the CTD of PTCH1. A recent study on PTCH1 revealed that simultaneous deletion of both its TM6–TM7 internal loop and its cytoplasmic domain did not affect PTCH1-dependent repression of SMO activity in PTCH1-deficient MEFs or normal localization in cilia[21]. This suggests a structural or mechanistic interaction between the TM6–TM7 internal loop and the CTD, because deletion of the CTD in combination with this loop restores normal activity. This PTCH1 variant (PTCH1*) has better solubility (Extended Data Fig. 2b). To test the function of PTCH1*, either PTCH1* or full-length PTCH1 was transfected to PTCH1-deficient MEFs. HH reporter assays reveal that, similar to wild-type PTCH1, PTCH1* can repress HH signalling. They also reveal that treatment with conditioned medium containing palmitoylated SHH-N without the cholesterol modification, but not the C24S palmitoylation-site SHH-N mutant, can release this repression (Fig. 1b).

We assembled the PTCH1*–SHH-N complex using unmodified SHH-N with an N-terminal His tag purified from *Escherichia coli*, or SHH-N with a C-terminal cholesterol and an N-terminal fatty-acid modification purified from HEK-293 cells (termed 'native' SHH-N hereafter) (Fig. 1c). The native SHH-N, but not unmodified SHH-N, formed a stable complex with PTCH1*, with SHH-N being detected at a 1:1 ratio (Fig. 1c). We also measured HH signalling activity in cells by adding either conditioned media or purified SHH-N proteins to SHH Light II cells that carry a Gli reporter plasmid. The results show that palmitoylated SHH-N, but not the C24S SHH-N mutant or N-His-tagged SHH-N, can stimulate HH signalling considerably (Fig. 1d). This suggests that PTCH1* is able to bind the native SHH-N, allowing us to make a physiological complex with purified PTCH1* in vitro (Extended Data Fig. 2c).

We determined the structure of PTCH1* to 3.5 Å resolution (Fig. 2a, Extended Data Figs. 3, 4, Extended Data Table 1). PTCH1*, a monomer in solution, measures 110 Å × 60 Å × 40 Å (Fig. 2a). The structure exhibits pseudosymmetry across the 12 transmembrane helices and features two homologous extracellular domains (Fig. 2a). The transmembrane domain of Patched has a similar topology to NPC1 and to prokaryotic RND transporters[7,19], one of which is AcrB (Fig. 2b, c). Previous studies have suggested that PTCH1 could form oligomers mediated by its CTD[22]; because this part of the molecule has been removed in our construct, we focus our discussion on monomeric PTCH1*.

Previous crosslinking studies have suggested that the SSD of NPC1 may bind a small ligand[23]. We have shown previously that a cavity in the SSD of NPC1 is large enough to accommodate a cholesterol molecule[7]. A corresponding pocket is observed in the SSD of PTCH1* (Fig. 2d, Extended Data Fig. 5). This hydrophobic pocket opens to the extracellular space and plasma membrane and measures roughly 20 Å × 10 Å × 10 Å (Fig. 2d). Remarkably, we observed a rod-shaped density in this pocket (Fig. 2d), which is distinct from detergent micelles and other noise on the basis of its local resolution. We speculate that the density might derive from an endogenous sterol derivative or another lipid. Structural comparison of PTCH1* and NPC1 reveals that the transmembrane helices of the N-terminal half of PTCH1* converge more closely than those in NPC1, potentially owing to an interaction with this unidentified ligand (Fig. 2d).

ECD-I and ECD-II overlap with each other with a root-mean-square deviation (r.m.s.d.) of 3.8 Å (Cα atoms) and resemble the middle and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

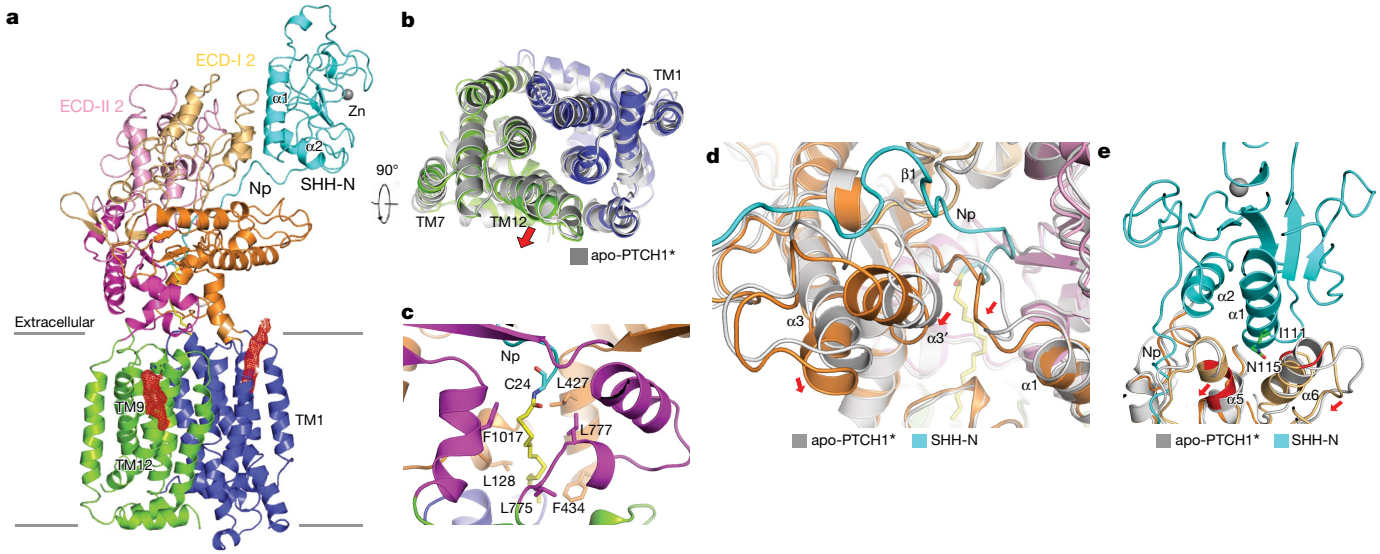


Fig. 3 | Structure of PTCH1*–SHH-N complex. a, Ribbon representation of the structure horizontal to the membrane with PTCH1*, coloured as in Fig. 2 and with SHH-N coloured in cyan. An SHH-N-bound zinc atom is indicated by a grey sphere. Putative endogenous molecules from the cryo-EM density map are shown at the 5σ level as a red mesh. **b**, Structural comparison of the membrane domains of apo-PTCH1* (grey) and PTCH1*–SHH-N (coloured). **c**, The palmitate-binding site. 'Np' denotes the N-terminal peptide of SHH-N (residues 24–38). **d**, Interface between the Np of SHH-N and ECD-I compared with apo-PTCH1*. **e**, Secondary interface between SHH-N and ECD-I subdomain 2 compared with apo-PTCH1*. Red arrows represent structural shifts.

C-terminal lumenal domains of NPC1[24]. Each ECD consists of two subdomains: subdomain 1 ranges from the cell membrane to the middle of each ECD, with three β strands and two α helices providing the main interface between the two extracellular domains; each subdomain 2 ranges from the middle to the top of each ECD (Fig. 2e). There are five residues of the swapped ECD-I α1 helix that interact with

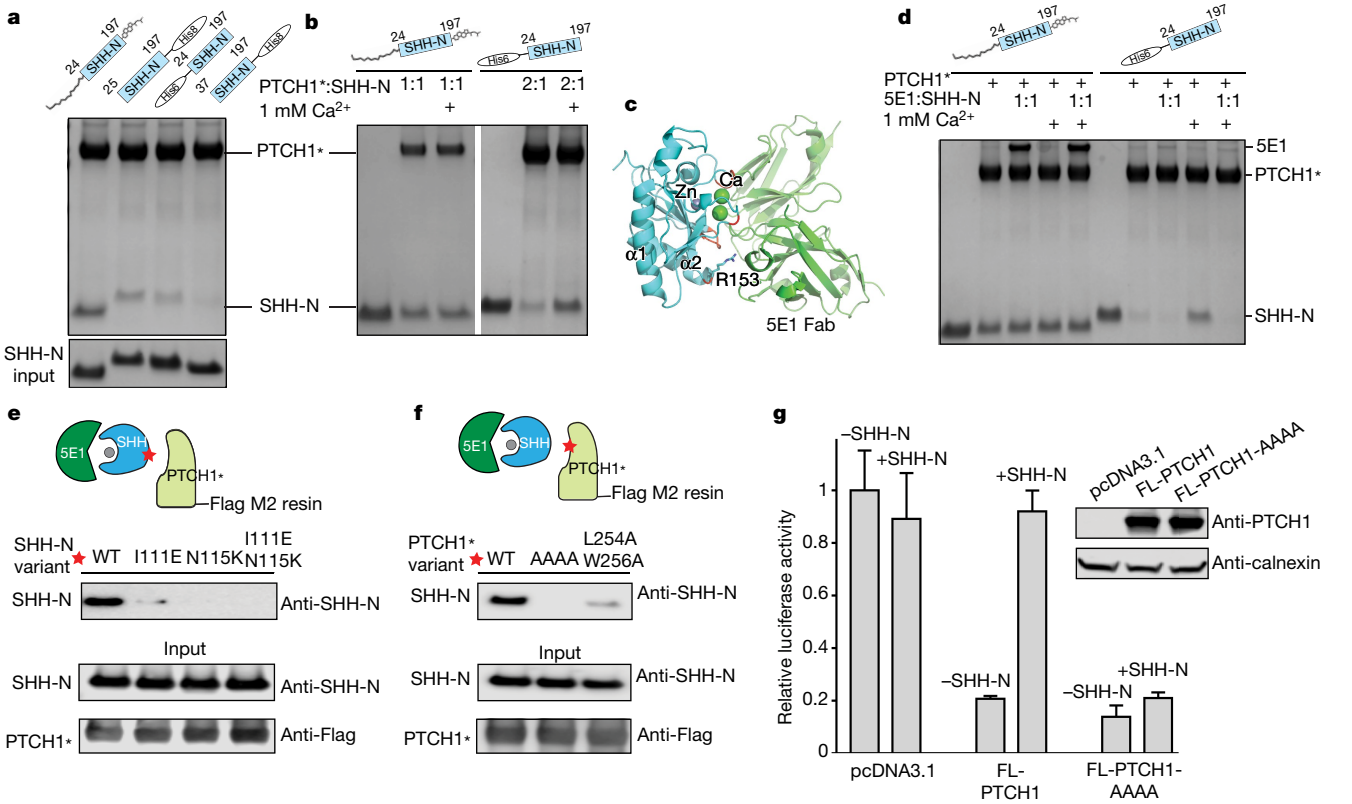


Fig. 4 | The palmitoylated N terminus of SHH-N dominates its interface to PTCH1*. a, The palmitoylated N terminus of SHH-N is important for PTCH1* binding. A 2:1 PTCH1*:SHH-N molar ratio was used for assays. **b**, Calcium facilitates the binding between PTCH1 and N-terminal His-tagged SHH-N. **c**, Complex of SHH-N and 5E1 (PDB code 3MXW), with interaction areas on SHH-N shown in red. Calcium atoms in the interface are indicated by green balls. **d**, 5E1 blocks the binding of PTCH1* to His-tagged SHH-N (right) but not to native SHH-N (left). **e**, Mutagenesis of the secondary interface of SHH-N. **f**, Mutagenesis of the interface of PTCH1*. The assays (**a**, **b**, **d**–**f**) were reproduced three times with similar results. **g**, Repression of HH signalling by PTCH1 with 'AAAA' mutations (PTCH1-AAAA) and its response to a SHH-N ligand. SHH-N in conditioned medium is shown in Fig. 1d. HH activity was measured using a luciferase assay and data are mean ± s.d. ($n = 3$). The protein was detected by Coomassie staining (**a**, **b**, **d**) or by western blotting (**e**–**g**). WT, wild-type.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

six residues in ECD-II in addition to several hydrophobic interactions. By contrast, ECD-II α1 (labelled as α1*) forms only one hydrophilic bond with T426 in α4 of ECD-I (Fig. 2f). Moreover, ECD-I contains 70 amino acids more than does ECD-II, and subdomain 2 of ECD-I contains more loops than does subdomain 2 of ECD-II (Extended Data Fig. 1). Together, these features confer a flexible character to ECD-I for ligand binding.

The structure of the PTCH1*–SHH-N complex was determined to 3.8 Å resolution (Fig. 3a, Extended Data Figs. 6, 7, Extended Data Table 1). The C terminus of SHH-N (residues 187–197), which was also invisible in the SHH-N crystal structure (Protein Data Bank (PDB) code 3M1N), could not be resolved. This result supports our cell biological assays that cholesterol modification of HH-N is not necessary to stimulate the HH signal (Fig. 1b, d). In addition to the molecule observed in the SSD, there is another endogenous density close to TM12 (Fig. 3a). This helix is slightly more tilted than in the apo-structure, potentially using the guanidine group of Arg1150 to bind the polar head of this putative lipid (Fig. 3b). There is no substantial conformational change between the transmembrane regions of PTCH1* alone and the PTCH1*–SHH-N complex (Fig. 3b). This suggests that SHH-N binding may not abolish SSD-mediated substrate binding of PTCH1.

Native SHH-N engages two binding sites on the ECD-I of PTCH1* (Fig. 3a). The primary interface involves the N-terminal peptide of SHH-N (residues 24–38), with fatty-acid modification that fits into the space between subdomains 1 of ECD-I and ECD-II (Fig. 3c). This binding site includes a strong stretch of density that extends beyond the N-terminal Cys24 and that is contiguous with the protein. Mass spectrometry has identified this extension on the SHH-N that we used as primarily palmitoylation and to a smaller degree lauryolation and myristoylation, consistent with previous observations[25]. On the basis of this result and the shape of the density, we assign the N-terminal density as a palmitoyl moiety (Fig. 3c, Extended Data Fig. 7e). Compared with the apo-PTCH1* structure, the α3 and α3′ helices and their connecting loop in ECD-I are shifted towards the membrane side, providing more space for this insertion (Fig. 3d). The loop that connects α1 and β1 of ECD-I (residues 148–153) undergoes a conformational change that allows palmitate to expand the space between α1 of ECD-I and α4 of ECD-II (Fig. 3d). Hydrophobic residues from extracellular domains form extensive van der Waals interactions with this modification (Fig. 3c). The secondary binding site involves α helices α1 and α2 of SHH-N, whereas PTCH1* engages its ECD-I (Fig. 3e). A recent study showed that a short palmitoylated N-terminal fragment (residues 24–45) of SHH-N could partially activate HH signalling by binding PTCH1[11]. Our work shows that native SHH-N indeed forms a more stable complex with PTCH1* than does unpalmitoylated SHH-N (Fig. 1c) and our structure confirms the interaction between the SHH-N palmitate and PTCH1* (Fig. 3c).

The interface that we observed between the native SHH-N and PTCH1* was inconsistent with a previously reported interface of SHH-N and PTCH1* that includes the zinc-binding site (Fig. 3a), which is also able to accommodate calcium and putatively binds to PTCH1[21,26,27]. To resolve this contradiction, we purified N-His-tagged SHH-N without the palmitate from *E. coli* as previously reported[26]; however, our cell-based HH reporter assays show that the N-His-tagged SHH-N lost almost all HH signalling activity (Fig. 1d). Compared with native SHH-N, removing the palmitate modification of SHH-N or deleting the N terminus (residues 24–36) weakened SHH-N binding to PTCH1* (Fig. 4a). Interestingly, binding between N-His-tagged SHH-N and PTCH1* could be enhanced by Ca^{2+}; however, the binding between native SHH-N and PTCH1* was not affected by Ca^{2+} (Fig. 4b).

We performed further SHH-N competition assays with PTCH1* and 5E1, a monoclonal anti-SHH-N antibody of nanomolar binding affinity that is used to block HH signalling by binding PTCH1[8] (Fig. 4c). Our structural analysis predicts that 5E1 should not interfere with the palmitate-dominated interface to PTCH1*. To validate this point, we performed a pull-down assay in the presence of 5E1 to

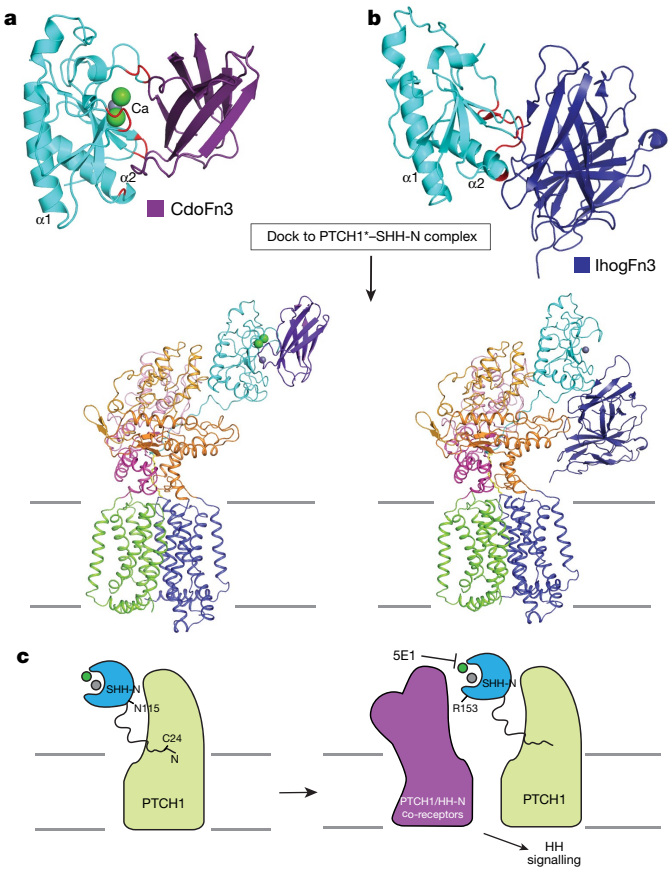


Fig. 5 | Putative multivalent complex of SHH-N with PTCH1 and its co-receptors. a, **b**, Complex of SHH-N and CdoFn3 (PDB code 3D1M) (**a**) and of SHH-N and IhogFn3 (PDB code 2IBG) (**b**), with interaction areas on SHH-N shown in red. Hypothetical models of the PTCH1*–SHH-N–Cdo complex or PTCH1*–SHH-N–Ihog complex (bottom panels) were generated by docking SHH-N–CdoFn3 or SHH-N–IhogFn3 to the PTCH1*–SHH-N structure. **c**, Model of the putative collaboration between SHH-N–PTCH1 and co-receptors or PTCH1 itself in HH signalling.

determine whether it competes with PTCH1* to bind native SHH-N or N-terminally tagged SHH-N. SHH-N was mixed with 5E1 before being incubated with PTCH1* immobilized on a Flag M2 resin. A PTCH1*–SHH-N–5E1 trimeric complex could be eluted, which suggests that binding of 5E1 to native SHH-N does not block its access to the observed PTCH1* interface. By contrast, 5E1 successfully competed with N-His-tagged SHH-N to bind PTCH1* with and without Ca^{2+} (Fig. 4d).

To exclude the possibility that detergent may have had an undesirable influence in our system, we also reconstituted PTCH1* with amphipols, which stabilize membrane proteins in solution, and repeated the competition assays (Extended Data Fig. 8a). The results show that, as in the presence of detergents, the PTCH1*–SHH-N–5E1 trimeric complex could be detected in a detergent-free environment (Extended Data Fig. 8b). Therefore, our data suggest that the palmitoylated N terminus of SHH-N is an integral part of the native SHH-N–PTCH1* interface and that when its palmitoylated N terminus is absent the 5E1-binding interface (including R153 and the Ca^{2+} binding site) may dominate in PTCH1* binding (Fig. 4c).

To verify the secondary interface further, we used unmodified SHH-N without any lipidation or tag at the N terminus and further blocked the previously reported interface on SHH-N with 5E1. In our binding assays, unmodified SHH-N still bound PTCH1* by the secondary interface in the presence of 5E1, which blocks the previously reported interface (Fig. 4e). We then introduced mutations on helix α1 of SHH-N(I111E/N115K), which abolished binding to PTCH1*

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

(Fig. 4e). The HH reporter assays show that the SHH-N(I111E/N115K) mutant-conditioned medium lost more than 70% of activity compared with wild-type SHH-N (Fig. 1d). These data may explain why mutation of I111 or N115 leads to HPE-3, possibly by altering how PTCH1 recognizes SHH-N. We also introduced alanine mutations on the ECD-I(E221A/Y222A/L223A/Y224A) of PTCH1* (shown as 'AAAA' and L254A/W256A in Fig. 3e in red). The binding assays show that these two mutants have weaker binding to the SHH-N–5E1 complex, further supporting our structural observations (Fig. 4f). The HH reporter assays also show that in *Ptch1*$^{-/-}$ MEFs the full-length PTCH1 with 'AAAA' mutations can repress HH signalling but cannot recognize SHH-N to release the inhibition (Fig. 4g).

Here, we have reported two structures and related structure-guided experiments that together reveal that PTCH1* recognizes native SHH-N by two distinct binding sites. The PTCH1*–SHH-N interface that we describe here has important implications of how SHH-N recognizes and interacts with other proteins in various signalling pathways, including Ihog (interference hedgehog), Cdo (cell-adhesion-molecule-related, downregulated by oncogenes), Boc (brother of Cdo) and Hhip (hedgehog-interacting protein)[8]. These co-receptors function in the recognition and localization of HH in various cell types[28,29]. Previous studies have suggested that HH-N could form a complex with Patched and its co-receptor[28,30]. The PTCH1*–SHH-N complex structure that we determined indeed allows for SHH-N to interact with an additional co-receptor to form multivalent complexes (Fig. 5a, b). This architecture is corroborated by our finding that the interaction of SHH-N with the antibody 5E1, which binds the same area of PTCH1 as does the SHH-N co-receptors, does not interfere with the formation of the PTCH1*–SHH-N complex (Fig. 4d). We propose a PTCH1–SHH-N working model: SHH-N initially recognizes PTCH1 through its palmitoylated N terminus; subsequently, SHH-N co-receptors or another PTCH1 binds SHH-N at a distinct interface to further regulate HH signalling (Fig. 5c). This model could provide a possible mechanism for how HH-N co-receptors and PTCH1 orchestrate HH signalling. The other aspect of Patched signalling—the mechanism for inhibition of SMO—remains poorly understood. Further investigations are required into how the HH-N ligand affects the putative transport activity of Patched.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0308-7.

Received: 25 October 2017; Accepted: 6 June 2018;
Published online 11 July 2018.

1. Ingham, P. W. & McMahon, A. P. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev.* **15**, 3059–3087 (2001).
2. Jiang, J. & Hui, C. C. Hedgehog signaling in development and cancer. *Dev. Cell* **15**, 801–812 (2008).
3. Briscoe, J. & Thérond, P. P. The mechanisms of Hedgehog signalling and its roles in development and disease. *Nat. Rev. Mol. Cell Biol.* **14**, 416–429 (2013).
4. Hooper, J. E. & Scott, M. P. Communicating with Hedgehogs. *Nat. Rev. Mol. Cell Biol.* **6**, 306–317 (2005).
5. Taipale, J. & Beachy, P. A. The Hedgehog and Wnt signalling pathways in cancer. *Nature* **411**, 349–354 (2001).
6. Rubin, L. L. & de Sauvage, F. J. Targeting the Hedgehog pathway in cancer. *Nat. Rev. Drug Discov.* **5**, 1026–1033 (2006).
7. Li, X. et al. Structure of human Niemann-Pick C1 protein. *Proc. Natl Acad. Sci. USA* **113**, 8212–8217 (2016).
8. Beachy, P. A., Hymowitz, S. G., Lazarus, R. A., Leahy, D. J. & Siebold, C. Interactions between Hedgehog proteins and their binding partners come into view. *Genes Dev.* **24**, 2001–2012 (2010).
9. Pepinsky, R. B. et al. Identification of a palmitic acid-modified form of human Sonic hedgehog. *J. Biol. Chem.* **273**, 14037–14045 (1998).
10. Chamoun, Z. et al. Skinny hedgehog, an acyltransferase required for palmitoylation and activity of the hedgehog signal. *Science* **293**, 2080–2084 (2001).
11. Tukachinsky, H., Petrov, K., Watanabe, M. & Salic, A. Mechanism of inhibition of the tumor suppressor Patched by Sonic Hedgehog. *Proc. Natl Acad. Sci. USA* **113**, E5866–E5875 (2016).
12. Williams, K. P. et al. Functional antagonists of sonic hedgehog reveal the importance of the N terminus for activity. *J. Cell Sci.* **112**, 4405–4414 (1999).
13. Kohtz, J. D. et al. N-terminal fatty-acylation of sonic hedgehog enhances the induction of rodent ventral forebrain neurons. *Development* **128**, 2351–2363 (2001).
14. Petrova, E., Rios-Esteves, J., Ouerfelli, O., Glickman, J. F. & Resh, M. D. Inhibitors of Hedgehog acyltransferase block Sonic Hedgehog signaling. *Nat. Chem. Biol.* **9**, 247–249 (2013).
15. Goldstein, J. L., DeBose-Boyd, R. A. & Brown, M. S. Protein sensors for membrane sterols. *Cell* **124**, 35–46 (2006).
16. Ingham, P. W., Taylor, A. M. & Nakano, Y. Role of the *Drosophila* patched gene in positional signalling. *Nature* **353**, 184–187 (1991).
17. Taipale, J., Cooper, M. K., Maiti, T. & Beachy, P. A. Patched acts catalytically to suppress the activity of Smoothened. *Nature* **418**, 892–896 (2002).
18. Ingham, P. W. et al. Patched represses the Hedgehog signalling pathway by promoting modification of the Smoothened protein. *Curr. Biol.* **10**, 1315–1318 (2000).
19. Gong, X. et al. Structural insights into the Niemann-Pick C1 (NPC1)-mediated cholesterol transfer and Ebola infection. *Cell* **165**, 1467–1478 (2016).
20. Sharpe, H. J., Wang, W., Hannoush, R. N. & de Sauvage, F. J. Regulation of the oncoprotein Smoothened by small molecules. *Nat. Chem. Biol.* **11**, 246–255 (2015).
21. Fleet, A., Lee, J. P., Tamachi, A., Javeed, I. & Hamel, P. A. Activities of the cytoplasmic domains of Patched-1 modulate but are not essential for the regulation of canonical Hedgehog signaling. *J. Biol. Chem.* **291**, 17557–17568 (2016).
22. Lu, X., Liu, S. & Kornberg, T. B. The C-terminal tail of the Hedgehog receptor Patched regulates both localization and turnover. *Genes Dev.* **20**, 2539–2551 (2006).
23. Lu, F. et al. Identification of NPC1 as the target of U18666A, an inhibitor of lysosomal cholesterol export and Ebola infection. *eLife* **4**, e12177 (2015).
24. Li, X. et al. 3.3 Å structure of Niemann-Pick C1 protein reveals insights into the function of the C-terminal luminal domain in cholesterol transport. *Proc. Natl Acad. Sci. USA* **114**, 9116–9121 (2017).
25. Long, J. et al. Identification of a family of fatty-acid-speciated sonic hedgehog proteins, whose members display differential biological properties. *Cell Rep.* **10**, 1280–1287 (2015).
26. McLellan, J. S. et al. The mode of Hedgehog binding to Ihog homologues is not conserved across different phyla. *Nature* **455**, 979–983 (2008).
27. Pepinsky, R. B. et al. Mapping sonic hedgehog-receptor interactions by steric interference. *J. Biol. Chem.* **275**, 10995–11001 (2000).
28. Allen, B. L. et al. Overlapping roles and collective requirement for the coreceptors GAS1, CDO, and BOC in SHH pathway function. *Dev. Cell* **20**, 775–787 (2011).
29. Yao, S., Lum, L. & Beachy, P. The ihog cell-surface proteins bind Hedgehog and mediate pathway activation. *Cell* **125**, 343–357 (2006).
30. Izzi, L. et al. Boc and Gas1 each form distinct Shh receptor complexes with Ptch1 and are required for Shh-mediated cell proliferation. *Dev. Cell* **20**, 788–801 (2011).

Acknowledgements We appreciate support and suggestions from Günter Blobel, and dedicate this manuscript to him. We thank M. Ebrahim and J. Sotiris at the Evelyn Gruss Lipper Cryo-EM Resource Center of the Rockefeller University for assistance with data collection and D. Nicastro and Z. Chen at the UT Southwestern Cryo-EM Facility (funded in part by the CPRIT Core Facility Support Award RP170644) for facility access and data acquisition; L. Beatty for help with tissue culture; A. Lemoff at the UT Southwestern Proteomics Core for mass spectrometry identification; B. Chen and J. Kim for 5E1 antibody, SHH Light II cells and *Ptch1*$^{-/-}$ MEFs; and M. Brown, E. Debler, J. Goldstein, J. Jiang, D. Rosenbaum and Z. Zhang for discussion. This work was supported by the Endowed Scholars Program in Medical Science of UT Southwestern Medical Center and O'Donnell Junior Faculty Funds (to X.L.), by NIH grant P01 HL020948 (Tissue Culture Core), by the Rockefeller University (to E.C.) and by the National Key Research and Development Program of MOST (numbers 2016YFA0501103 and 2015CB910104 to J.W.). X.L. is the Rita C. and William P. Clements, Jr. Scholar in Biomedical Research of UT Southwestern.

Reviewer information *Nature* thanks F. de Sauvage and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions X.L. conceived the project. X.Q., P.S. and X.L. purified the protein and performed the functional characterization. X.Q., P.S., E.C. and X.L. carried out cryo-EM work. J.W. built the initial model and refined the structures with X.Q. and X.L. All the authors designed the research, analysed the data and contributed to manuscript preparation.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0308-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0308-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.W. or X.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Protein expression and purification. The constructs of human patched 1 were cloned into pEG BacMam with a C-terminal Flag tag. The protein was expressed using baculovirus-mediated transduction of mammalian HEK-293S GnTI$^-$ cells (ATCC). The cell lines tested negative for mycoplasma contamination. At 48 h post-infection at 37 °C, cells were disrupted by sonication in buffer A (20 mM Hepes pH 7.5, 150 mM NaCl) with 1 mM phenylmethanesulfonyl fluoride (PMSF), 10 μg ml^{-1} leupeptin. After low-speed centrifugation, the resulting supernatant was incubated in buffer A with 1% (w/v) *n*-dodecyl-β-D-maltoside (DDM, Anatrace) for 1 h at 4 °C. The lysate was centrifuged again and the supernatant was loaded onto an anti-Flag M2 affinity column (Sigma). After washing twice, the protein was eluted in buffer A with 0.1 mg ml^{-1} Flag peptide, 0.02% DDM, and concentrated. The concentrated protein was purified by Superdex-200 size-exclusion chromatography (GE Healthcare) in buffer B (20 mM Hepes pH 7.5, 150 mM NaCl and 0.06% (w/v) Digitonin (Sigma)). The peak fractions were collected and concentrated to 5–7 mg ml^{-1} for grid preparation. Mass spectrometry and anti-Flag-tag western blotting confirmed the identity of the protein. To assemble the PTCH1*–SHH-N complex, native SHH-N (purchased from R&D system, catalogue number 8908-SH/CF) was mixed with purified PTCH1* at a 1:1 molar ratio and purified by Superdex-200 size-exclusion chromatography (GE Healthcare) in buffer B. The peak fractions were collected and concentrated to 5–7 mg ml^{-1} for grid preparation.

For preparation of detergent-free protein, PTCH1* was purified as above and then mixed with Amphipol A8-35 (Anatrace) at a 1:3 mass ratio for 4 h. This mixture was incubated with Bio-Beads (Bio-Rad) overnight before further purification by gel filtration with buffer A. The mutated and truncated DNA constructs were generated using two-step PCR or Gibson assembly (NEB).

Three constructs were cloned into the pET21b vector: (1) human SHH-N (residues 24–197) with N-terminal His tag; (2) human SHH-N (residues 25–197) with C-terminal His tag; and (3) human SHH-N (residues 37–197) with C-terminal His tag. All of the constructs and SHH-N variants were then transformed into *E. coli* BL21 (DE3) for expression. The transformed bacteria were grown in LB medium supplemented with ampicillin at 37 °C and induced by 0.2 mM isopropyl β-D-thiogalactopyranoside (IPTG) overnight at 25 °C. The cells were harvested and lysed by sonication in buffer C (20 mM Hepes pH 7.5, 500 mM NaCl) supplemented with 1 mM PMSF. The lysate was centrifuged at 18,000 r.p.m. for 30 min and the supernatant was loaded onto a Ni^{2+}-NTA affinity column (Qiagen). After washing three times with 20 mM, 40 mM and 80 mM imidazole in buffer C, the protein was eluted in buffer C plus 250 mM imidazole and further purified by gel filtration using Superdex-200 size-exclusion chromatography (GE Healthcare) in buffer C. Peak fractions were collected for pull-down assay.

Electron microscopy sample preparation and imaging. A freshly purified protein sample was added to Quantifoil R1.2/1.3 400 mesh Au holey carbon grids (Quantifoil), blotted using a Vitrobot Mark IV (FEI), and frozen in liquid ethane. The grids were imaged in a 300-keV Titan Krios (FEI) with a Gatan K2 Summit direct electron detector (Gatan). Data were collected at 1 Å per pixel with a dose rate of 8 electrons per physical pixel per second. Images were recorded for 10-s exposure in 50 subframes to give a total dose of 80 electrons per Å^2.

Imaging processing and 3D reconstruction. Dark subtracted images were first normalized by gain reference that resulted in a pixel size of 1 Å per pixel. Drift correction was performed using the program Unblur[31]. The contrast transfer function (CTF) was estimated using CTFFIND4[32]. To generate PTCH1* templates for automatic picking, around 2,000 particles were manually picked and classified by 2D classification in RELION[33]. After auto-picking in RELION, the low-quality images and false-positive particles were removed manually. About 790,000 particles were extracted for subsequent 2D and 3D classification. 3D classification was carried out in RELION to generate the initial model of PTCH1*, using the cryo-EM structure of human NPC1 (Electron Microscopy Data Bank, EMD-6640) low-pass-filtered to 60 Å as the initial model. The PTCH1* model of best class after 3D classification was used as the initial model for the final 3D classification in RELION. The best class, containing around 168,000 particles, provided a 7.7 Å map after 3D auto-refinement in RELION. Motion correction of all particles was performed using the program *alignparts_lmbfgs*[34]. As in a previously published approach[35], refinement was performed in FREALIGN[36] using this best class as the initial model. The global search was performed once followed by 10–20 rounds of local search without mask. The best class without mask refinement was selected to generate the mask using *relion_mask_create* with 6 Å extensions excluding the micelle. This mask was then used to perform another global search followed by 10–20 rounds of local search with a cosine edge width of 6 Å and a BSC value of 10 to exclude bad particles. The final map is estimated to be 3.5 Å using the 0.143 cut-off criteria.

To generate templates of the PTCH1*–SHH-N complex for automatic picking, around 5,000 particles were picked manually and classified by 2D classification in RELION[33]. After auto-picking in RELION, the low-quality images and false-positive particles were removed manually. About 661,000 particles were extracted for subsequent 2D and 3D classification. 3D classification was carried out in RELION using the cryo-EM structure of PTCH1* low-pass-filtered to 60 Å as the initial model. The complex model of best class after 3D classification was used as the initial model for the final 3D classification in RELION. The best class, containing 195,000 particles, provided a 7.1 Å map after 3D auto-refinement in RELION. After motion correction of individual particles, the final refinement was performed in FREALIGN[36] using this best class as the initial model. The global search was performed once followed by 10–20 rounds of local search without mask. The best class without mask refinement was selected to generate the mask using *relion_mask_create* with 6 Å extensions excluding the micelle. This mask was then used for performing another global search followed by 10–20 rounds of local search with a cosine edge width of 6 Å. The final map is estimated to be 3.8 Å using the 0.143 cut-off criteria.

Model construction. To obtain better side-chain densities for model building, we sharpened the map of PTCH1* using bfactor.exe (author, Nikolaus Grigorieff) with a resolution limit of 3.5 Å and a *B*-factor value of −100 Å^2. The entire model was built de novo in Coot[37]. The crystal structure of human NPC1 (residues 334–1,278, PDB code 5U74) and the glycosylation sites of the PTCH1* extracellular domains were used to check the registers of our model. The de novo models of ideal helices were first put into the transmembrane region. Using the bulky size of some large side chains in the cryo-EM map, we assigned a sequence to the initial model. The model was refined using *phenix.real_space_refine* with real-space restraints, including secondary-structure, stereochemical, Ramachandran and rotamer restrains, to accommodate the bending of helices, maintain the stereochemistry of the helical structure and best fit the model and cryo-EM map. Finally, the two extracellular domains were added to the model gradually, at the same time as the sequence assignment. The density of residues 1–75 (N-terminal domain), 608–618 and 721–730 (TM6–TM7 linker), 888–901 (in ECD-II) and 1,177–1,188 (C terminus) is not resolved or built. Residues 191–198, 211–263, 379–391, 457–466, 864–887, 902–915 and 955–960 were built with polyalanine owing to limited local resolution. For the complex, we sharpened the map using bfactor.exe with a resolution limit of 3.8 Å and a *B*-factor value of −100 Å^2. The residues 149–153 of PTCH1* were also built with polyalanine in the complex. Using the EMfit docking program (developed by Michael G. Rossmann), a single solution was obtained for the orientation of SHH-N in the cryo-EM map. Structures of SHH-N (PDB code 3M1N) and PTCH1* were docked into our final cryo-EM maps of the complex in Coot[37].

Model refinement and validation. The models of PTCH1* and its complex with SHH-N were refined in real space using PHENIX[38] and in reciprocal space using Refmac with secondary-structure restraints and stereochemical restraints[39,40]. Structure factors were calculated from a half-map (working) using the program SFall[41]. Fourier shell correlations (FSCs) were calculated between the two half maps, the model against the working map, the other (free) half map and full (sum) map[42]. Local resolutions were estimated using Blocres[43]. MolProbity[44] was used to validate the geometries of the model. Structure figures were generated using PyMOL (http://www.pymol.org) and Chimera[45].

Pull-down assay. The unmodified human SHH-N proteins were expressed and purified from *E. coli* as described above. The HEK-293-derived native human SHH-N protein was purchased from R&D Systems (catalogue number 8908-SH/CF; see the LC/ESI-mass spectrometry analysis of this protein at http://bit.ly/2AohYCG). For the pull-down assay, purified PTCH1* protein was immobilized to 20 μl anti-Flag M2 resin, which was further incubated with unmodified or native SHH-N for 1 h at 4 °C in 150 μl buffer B. Then the resin was spun down and washed three times with buffer B. The protein complex was eluted with 20 μl buffer B supplemented with 0.3 mg ml^{-1} Flag peptide. 15 μl of the elution was loaded on SDS–PAGE for detection. To see whether 5E1 competes with PTCH1* when binding SHH-N, 5E1 (from B. Chen and J. Kim, UT Southwestern) was added to SHH-N before incubating with the PTCH1*-immobilized anti-Flag M2 resin. For the pull down of the SHH-N–5E1 complex, the SHH-N (*E. coli* expressed with C-terminal His tag) was incubated with 5E1 at 1:1 molar ratio, and then the pull-down assays were performed as above. The SHH-N protein and PTCH1* were detected by anti-SHH antibody (sc-365112, Santa Cruz Biotechnology) and anti-Flag antibody (M185, MBL Life Science). For the detergent-free assay, buffer B was replaced by buffer A. Each assay was reproduced at least three times.

HH reporter assays. Human SHH-N (24–197) was constructed into pcDNA3.1 vector with the signal sequence of human calreticulin at the N terminus as described previously[11]. Secreted SHH-N was produced in HEK-293 cells (ATCC) by transient transfection for 72 h and was collected in DMEM, with 0.5% fetal bovine serum (FBS). SHH Light II cells, a stable cell line expressing firefly luciferase with a 8X-Gli promoter and *Renilla* luciferase with a constitutive promoter (from B. Chen and J. Kim), were used to measure HH pathway activity. SHH Light II cells were treated with the conditioned medium or purified protein diluted in fresh DMEM with 0.5% newborn calf serum for 30 h. To measure the activity of PTCH1 variants in HH signalling, the 8X-Gli luciferase firefly reporter transgene, a constitutive *Renilla* luciferase transgene and a pcDNA3.1 vector encoding

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

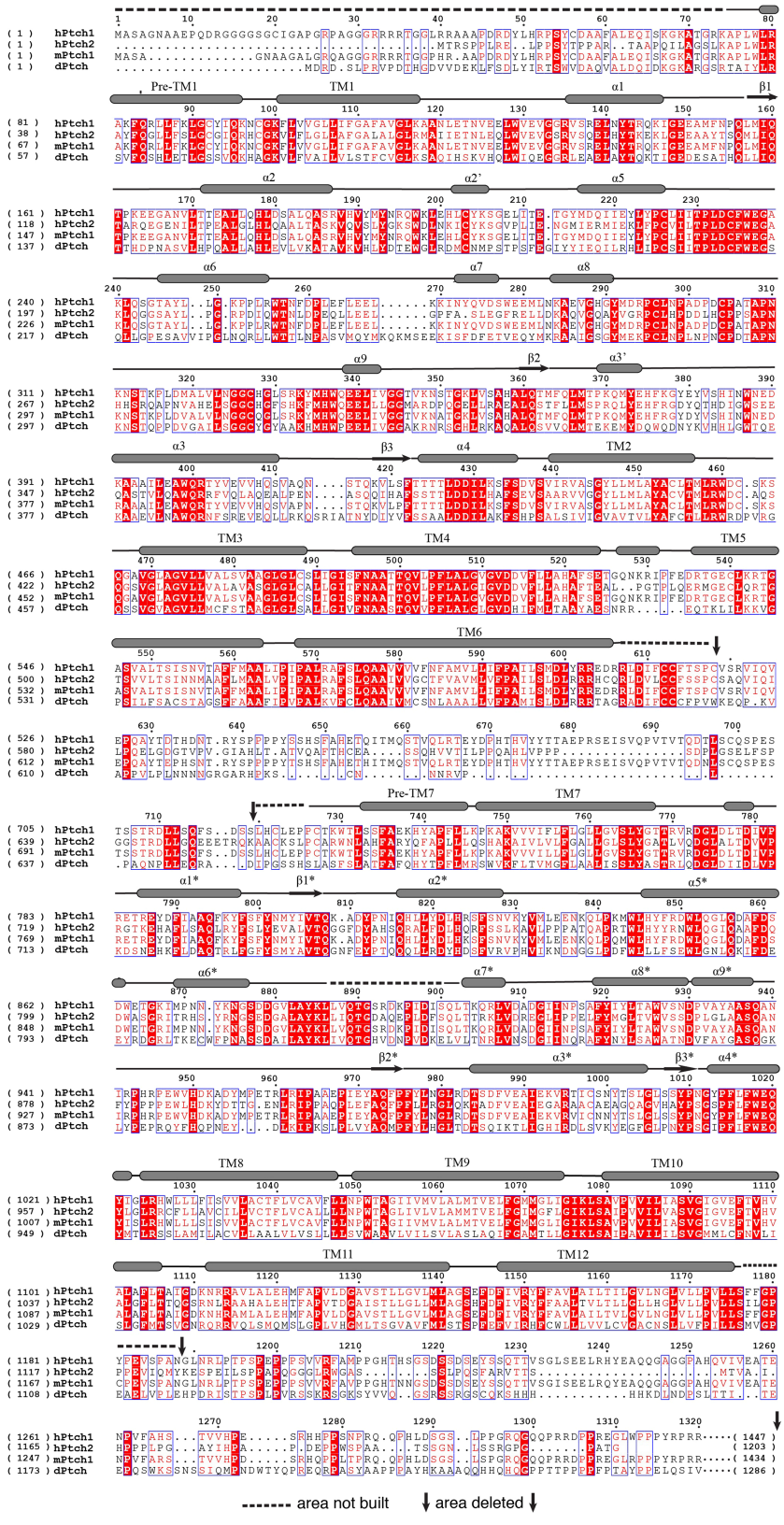
PTCH1 variants were transfected to *Ptch1*$^{-/-}$ MEFs (from B. Chen and J. Kim) using TransIT reagent (Mirus Bio LLC). After 24 h, cells were serum-starved in DMEM with 0.5% FBS. 24 h later, cells were treated with SHH-N conditioned medium for another 24 h. Firefly and *Renilla* luciferase were measured using the Dual-Luciferase Reporter Assay System (Promega). The conditioned medium added was normalized on the basis of western blotting with anti-SHH antibody. The expression of PTCH1 variants and internal calnexin in MEFs were detected by western blotting with anti-PTCH1 antibody (GeneTex, 83771) and anti-calnexin antibody (Novus, NB100-1965). Each assay was reproduced at least three times and data were analysed using Excel (Microsoft). Bar graphs were generated by Prism (GraphPad). The cell lines tested negative for mycoplasma contamination.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The 3D cryo-EM density maps of PTCH1* and PTCH1*–SHH-N have been deposited in the Electron Microscopy Data Bank under accession numbers EMD-7795 and EMD-7796. Atomic coordinates for the atomic model of PTCH1* and the PTCH1*–SHH-N complex have been deposited in the Protein Data Bank under accession numbers 6D4H and 6D4J. All other data are available from the corresponding authors on reasonable request.

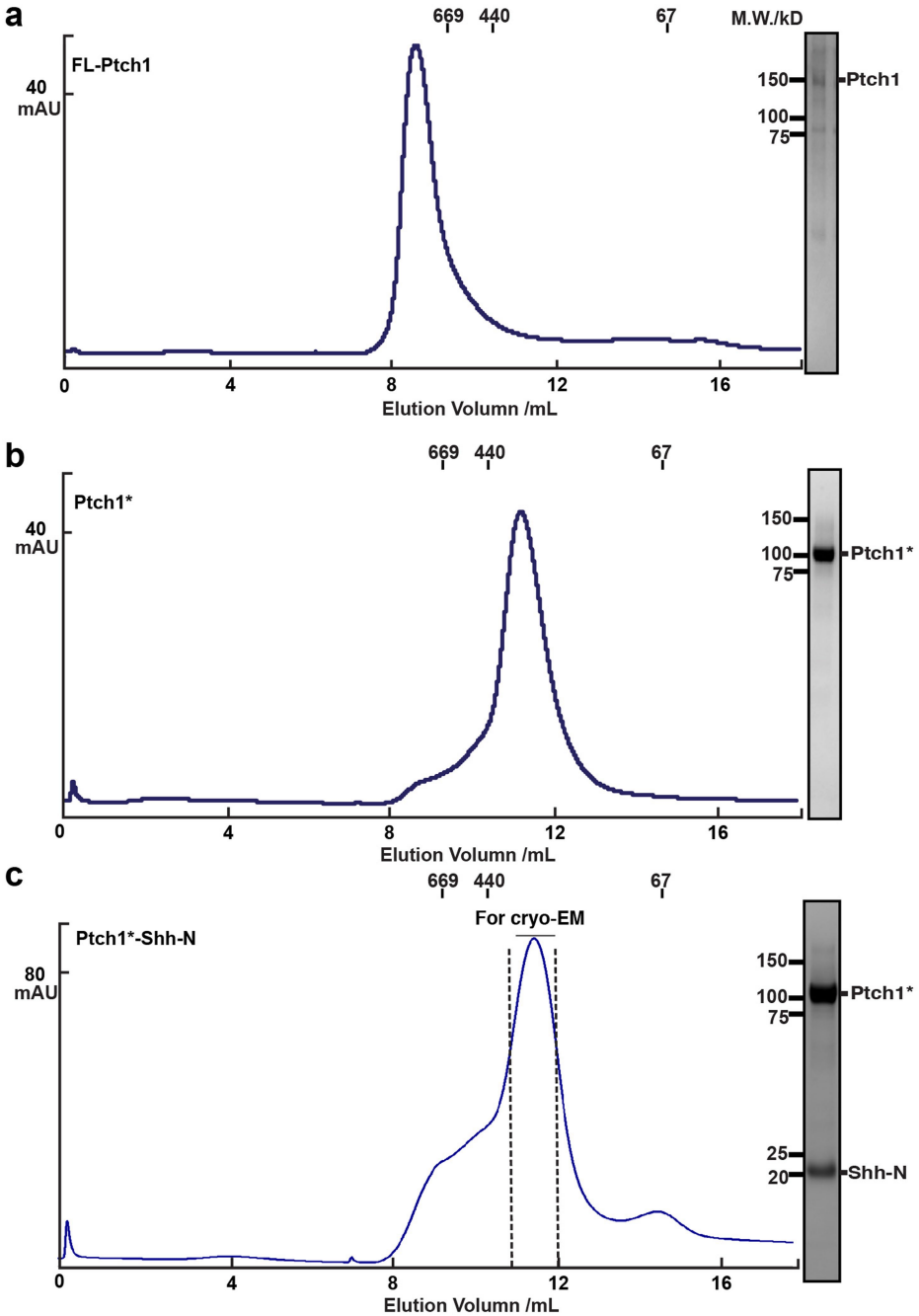
31. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4**, e06980 (2015).
32. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol*. **192**, 216–221 (2015).
33. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol*. **180**, 519–530 (2012).
34. Rubinstein, J. L. & Brubaker, M. A. Alignment of cryo-EM movies of individual particles by optimization of image translations. *J. Struct. Biol*. **192**, 188–195 (2015).
35. Zhang, Z., Liu, F. & Chen, J. Conformational changes of CFTR upon phosphorylation and ATP binding. *Cell* **170**, 483–491, e488 (2017).
36. Grigorieff, N. Frealign: an exploratory tool for single-particle cryo-EM. *Methods Enzymol*. **579**, 191–226 (2016).
37. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
38. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
39. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
40. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71**, 136–153 (2015).
41. Ten Eyck, L. F. Efficient structure-factor calculation for large molecules by the fast fourier transform. *Acta Crystallogr. A* **33**, 486–492 (1977).
42. Wang, Z. et al. An atomic model of brome mosaic virus using direct electron detection and real-space optimization. *Nat. Commun*. **5**, 4808 (2014).
43. Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol*. **157**, 3–18 (2007).
44. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
45. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem*. **25**, 1605–1612 (2004).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



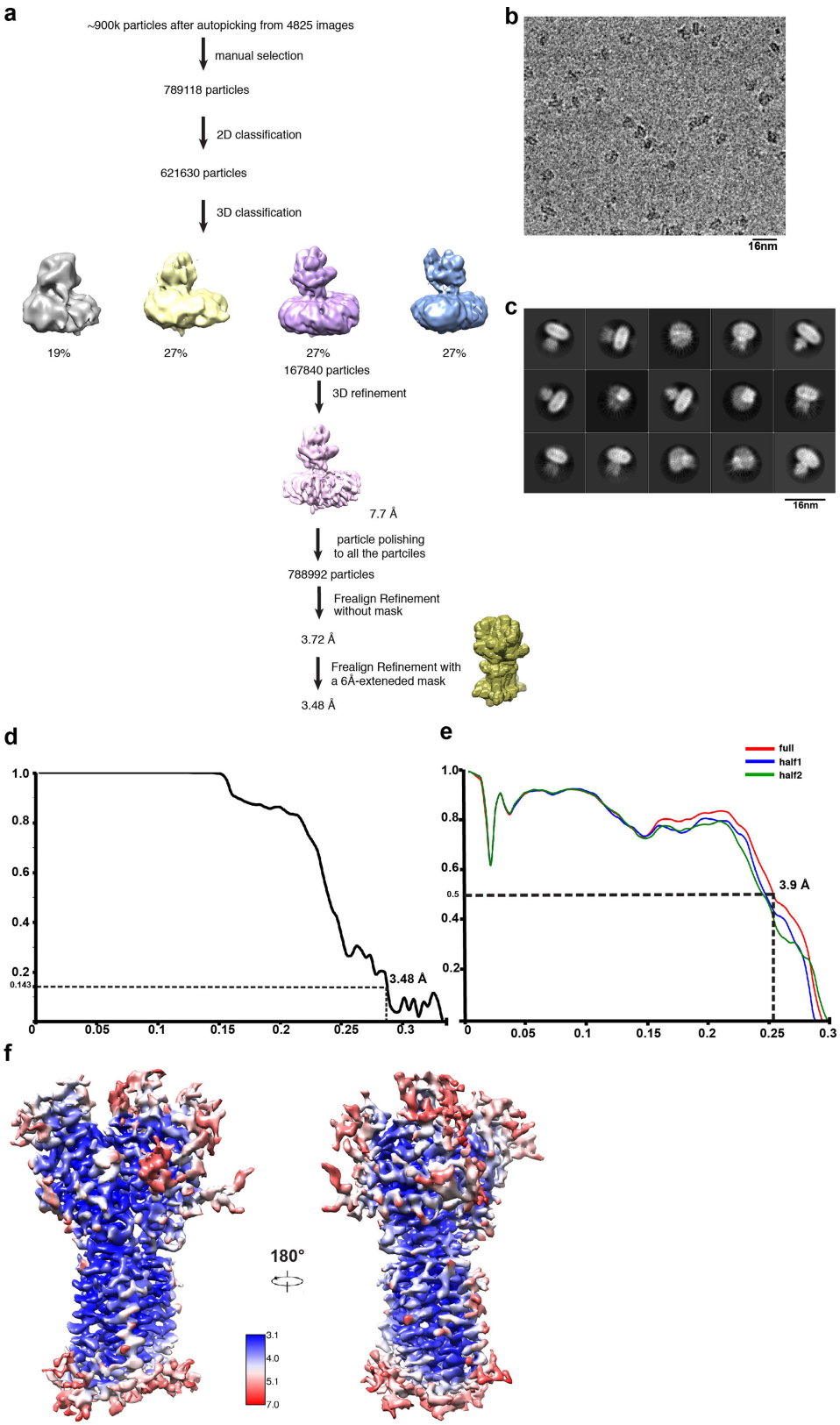
Extended Data Fig. 1 | Sequence alignment of human PTCH1 and PTCH2, mouse PTCH1 and *Drosophila* Ptc. The residue numbers of human PTCH1 are indicated above the protein sequence. The transmembrane helices and secondary structures of extracellular domains are labelled (structural elements of ECD-II with asterisk). Residues under the dashed lines are excluded from the 3D reconstruction.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



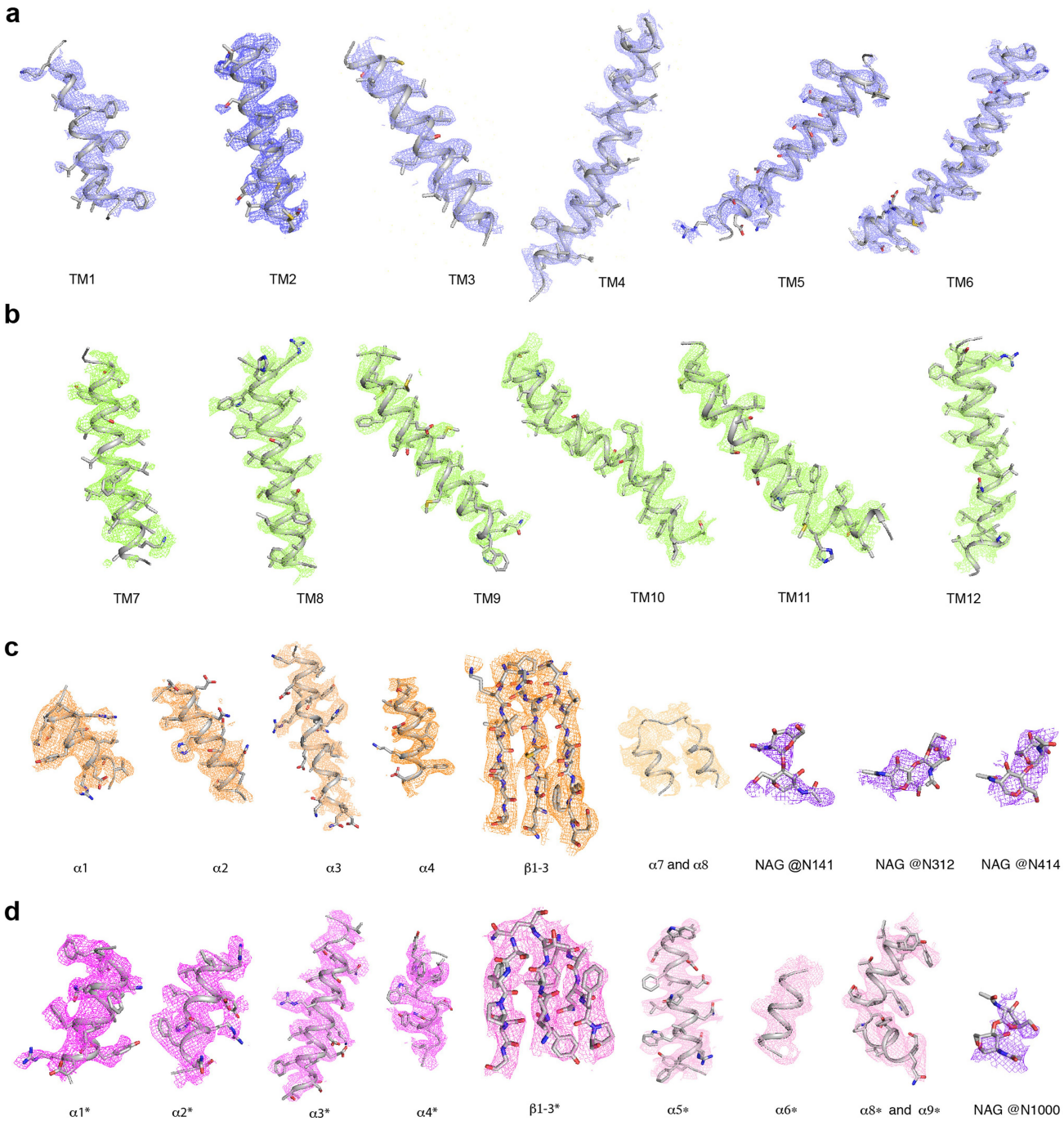
Extended Data Fig. 2 | Biochemical properties of expressed human PTCH1 proteins. a–c, Size-exclusion chromatogram and SDS–PAGE gel of the purified full-length PTCH1 (**a**), the purified PTCH1* (**b**) and the purified PTCH1*–SHH-N complex (**c**). Molecular standards are indicated on the left side of the gels and above the elution curves. The assays were reproduced at least three times with similar results.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



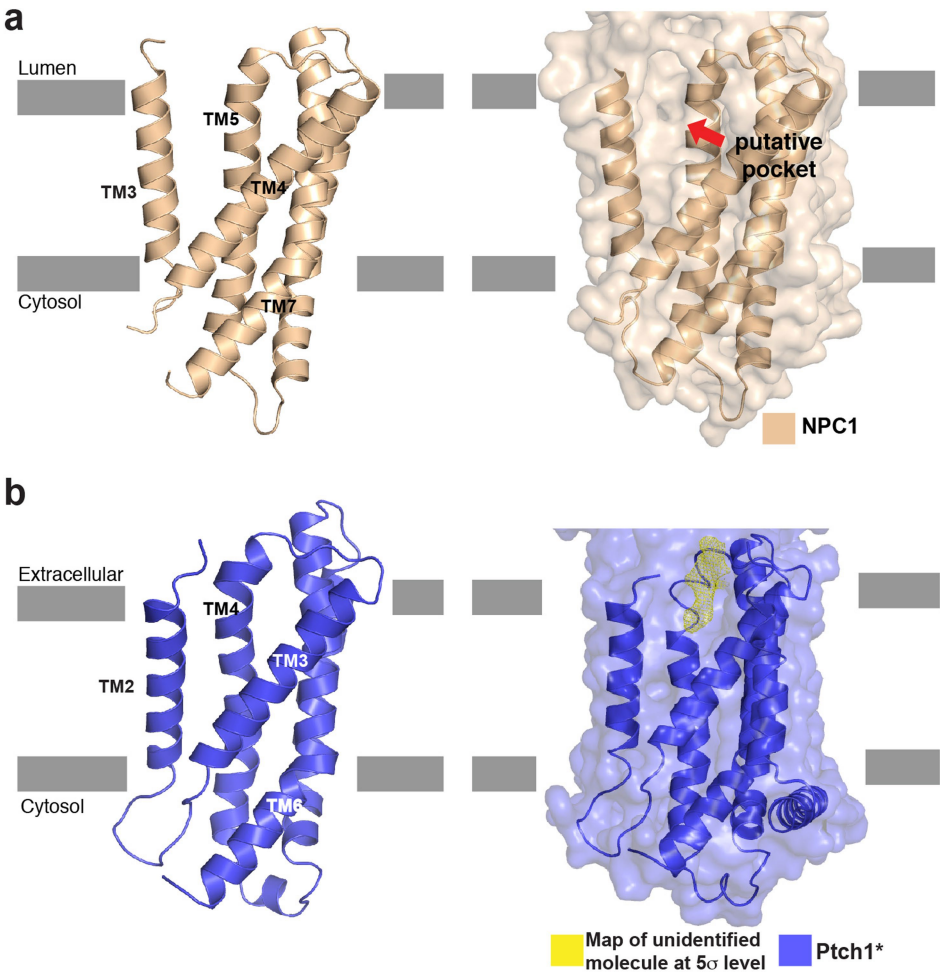
Extended Data Fig. 3 | Data processing and model quality assessment of PTCH1*. **a**, The data processing workflow for PTCH1*. **b**, A representative electron micrograph at a defocus of −2.0 μm. **c**, 2D classification. **d**, FSC curve of the structure as a function of resolution using Frealign output. **e**, The FSC curves calculated between the refined structure and the half map used for refinement, the other half map and the full map. **f**, Density maps of PTCH1* structure coloured by local resolution estimate using Blocres.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



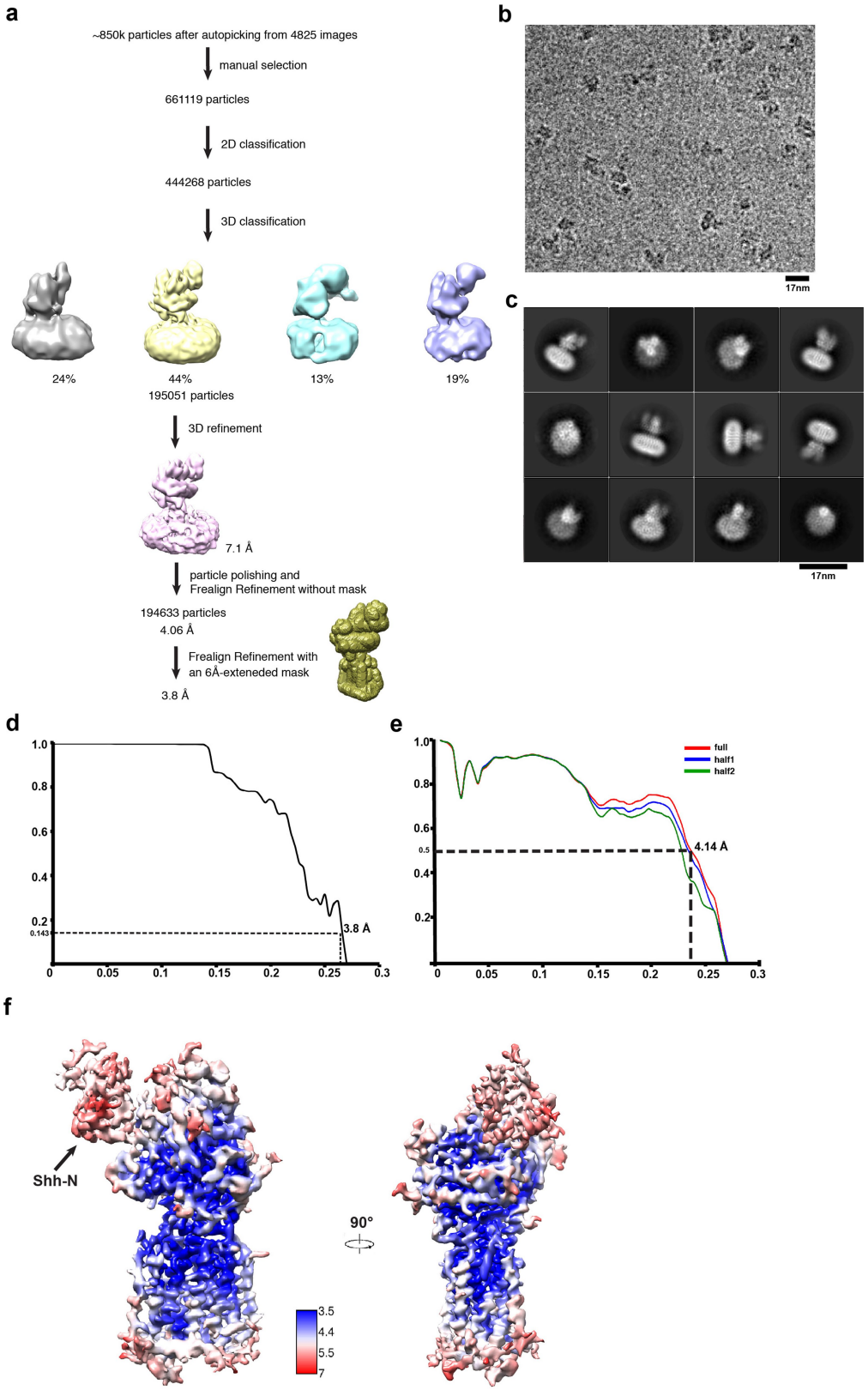
Extended Data Fig. 4 | Electron microscopy density of different portions of PTCH1* at the 5σ level. a, TM1–TM6. **b**, TM7–TM12. **c**, ECD-I. **d**, ECD-II. NAG, *N*-acetylglucosamine.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



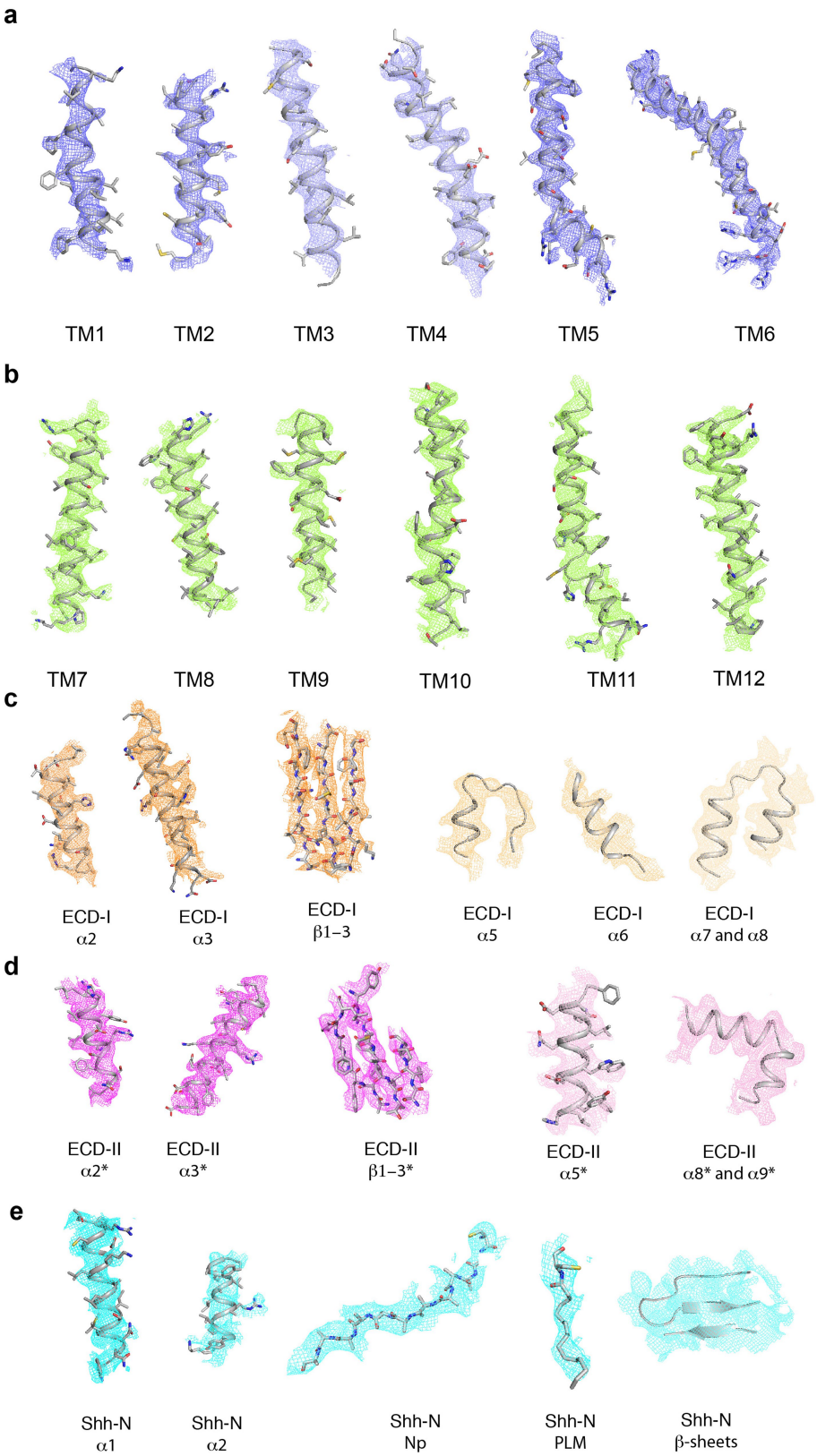
Extended Data Fig. 5 | NPC1 and PTCH1* SSD structural and surface comparison. a, NPC1 SSD. The putative pocket (indicated by the red arrow) in the SSD is created by TM3–TM5. **b**, PTCH1* SSD.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



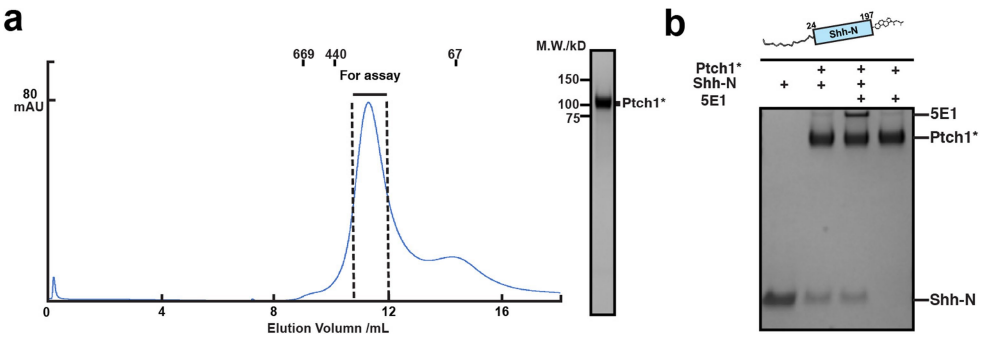
Extended Data Fig. 6 | Data processing and model quality assessment of PTCH1*–SHH-N. a–f, Same as Extended Data Fig. 3 but for PTCH1*–SHH-N.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Electron microscopy density of different portions of PTCH1*–SHH-N complex. a, TM1–TM6 at the 5σ level. **b**, TM7–TM12 at the 5σ level. **c**, Major structural elements of ECD-I at the 4.5σ level, **d**, Major structural elements of ECD-II at the 4.5σ level. **e**, Major structural elements of SHH-N at the 4.5σ level; palmitate (PLM) at the 3σ level.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | PTCH1*–SHH-N binding assay in the detergent-free system. a, Size-exclusion chromatogram and SDS–PAGE gel of the purified PTCH1* with Amphipol A8-35 in buffer A. Molecular standards are indicated on the left side of the gels and above the elution curves. **b**, 5E1 does not compete with the binding of native SHH-N to PTCH1*. 5E1 and SHH-N at a 1:1 molar ratio were incubated with PTCH1*-immobilized Flag M2 resin; the complex was eluted by Flag peptide. Protein was detected by Coomassie staining. The assay was reproduced three times with similar results.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	Ptch1* (EMDB-7795) (PDB 6D4H)	Ptch1*-Shh-N (EMDB-7796) (PDB 6D4J)
Data collection and processing		
Magnification	29000	29000
Voltage (kV)	300	300
Electron exposure (e–/Å^2)	80	80
Defocus range (μm)	-0.8 to -2.0	-0.8 to -2.2
Pixel size (Å)	1.0	1.0
Symmetry imposed	C1	C1
Initial particle images (no.)	789,118	661,119
Final particle images (no.)	167,840	195,051
Map resolution (Å)	3.48	3.80
FSC threshold 0.143		
Map resolution range (Å)	3.1-7.0	3.5-7.0
Refinement		
Initial model used (PDB code)	3JD8	6D4H
Model resolution (Å)	3.90	4.14
FSC threshold 0.5		
Model resolution range (Å)	3.90-256	4.14-256
Map sharpening *B* factor (Å^2)	-100	-100
Model composition		
Non-hydrogen atoms	7318	8614
Protein residues	964	1129
Ligands	8	8
B factors (Å^2)		
Protein	127.8	184.9
Ligand	177.2	218.2
R.m.s. deviations		
Bond lengths (Å)	0.0069	0.0063
Bond angles (°)	1.1545	1.0467
Validation		
MolProbity score	2.09	2.05
Clashscore	4.21	2.89
Poor rotamers (%)	3.12	3.69
Ramachandran plot		
Favored (%)	91.2	90.2
Allowed (%)	8.8	9.8
Disallowed (%)	0	0

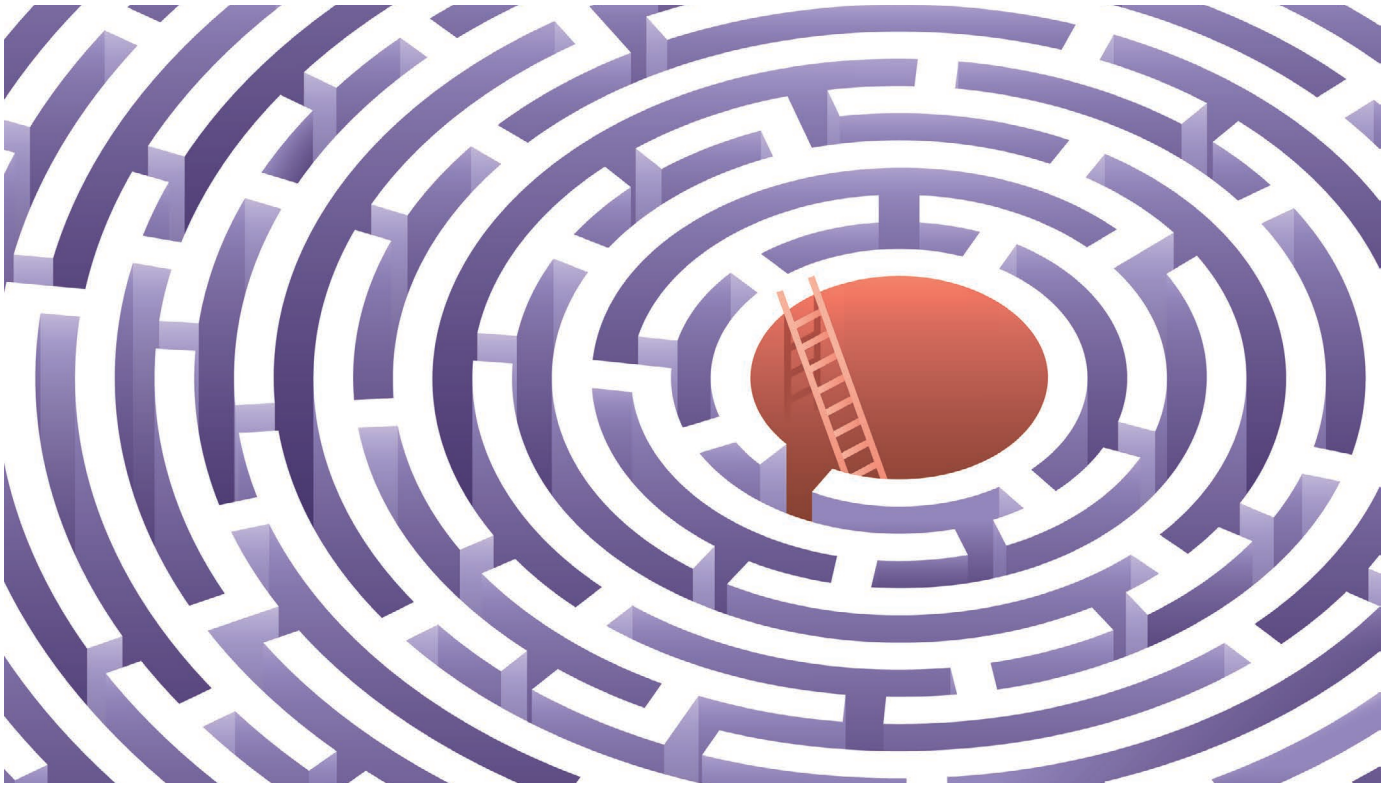
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

TWITTER Career news, advice and science jobs **https://twitter.com/NatureJobs**

BLOG Personal stories and careers counsel **blogs.nature.com/naturejobs**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

ADAPTED FROM GETTY



COLUMN

You are not a failed scientist

PhD students who leave academia should be supported, says **Philipp Kruger**.

As a PhD student in my final year, I find it demoralizing and frustrating to be constantly reminded of the bleak job prospects in academia. This dim outlook may well increase the pressure on students and contribute to high rates of anxiety and depression among them (T. M. Evans *et al. Nature Biotechnol.* **36**, 282–284; 2018). The scientific community could help to solve this problem by encouraging us all to change the way we think about the PhD. And scientists can start by appreciating a simple truth: researchers who leave academia are not failed academics.

Students and their supervisors must begin to regard a PhD programme as a traineeship in scientific thinking and an invaluable qualification for a diverse range of careers. If everyone involved in academic science could accept a variety of roles as the default outcome, we could change our flawed definition of success. We could transition from a culture of failure to a healthier and happier scientific enterprise.

I've found it daunting to determine the best career to match my personality, skills, priorities, ambitions and interests, particularly because most people around me treat the academic path as the default. But we PhD students have an obligation to see to our own professional futures. Many of us, of course, are driven by the excitement of discovery, and we relish the freedom to pursue our curiosity in an academic laboratory.

However, some of us discover during our PhDs that in our dream job, the emphasis would be on using interpersonal and communication skills, having a more immediate impact on society or gaining financial rewards, job security or family-friendly working hours. Our direction should be the result of a conscious decision rather than a perception of a lack of opportunities. And it should have nothing to do with a sense or fear of 'failure'.

WHAT PHD STUDENTS CAN DO

To make an informed decision about their professional future, students need to use their initiative and be proactive. Here's what has worked for me and what I think my student peers could do.

Determine your preferences. A PhD provides you with diverse experiences, from conducting lab experiments to giving talks and supervising students. I've found it useful to write down whenever one of these tasks is particularly fascinating, enjoyable, stressful, boring or frustrating. Doing this over four years has ▶

© 2018 Springer Nature Limited. All rights reserved.

helped me to determine patterns and decide which tasks should make up my future job. For example, this is how I discovered that I enjoy discussing and presenting my data much more than actually collecting them in the lab.

Develop a range of skills. It was important for me to be aware of and actively develop transferable skills, such as teamwork, line management of more-junior students, project and time management, written communication and conference presentations, from an early point in my PhD. You could seek activities outside the lab that provide further experience, such as organizing events, teaching, public engagement or writing a blog. I have, for example, co-organized conferences in Oxford for the British Society for Immunology, and I'm teaching immunology in various undergraduate programmes at the University of Oxford, UK.

Find out what's out there. This is not as straightforward as it should be, but I found that my university's careers service was a good place to start. Career advisers regularly encounter PhD students with doubts about their future roles, and they can help you to structure your self-assessment and broadly match your preferences and skills to specific roles. I also read job adverts to find out about available positions and the skills that employers seek. Another useful resource is science-careers fairs, where you can meet potential employers. Even after a lot of research online, I encountered jobs at those fairs that I had never heard of.

Organize talks. Don't depend solely on online articles and blog posts to understand what a certain role is really like. For me, the most substantial insights have come from talking to people who actually have those jobs. I've found that those who have decided to leave the lab for other professions know how difficult it is and are willing to give talks about their journey. Initiation of such careers seminars often comes from students, and you could even organize one yourself. In my department, we plan four of those talks each year. Send this article to your department or programme director when you ask them to fund your event.

Build your network. Early on, I found that my LinkedIn profile was a helpful tool for keeping track of connections and providing information about myself. If you send follow-up messages to people you have spoken to at seminars, meetings or fairs, they are more likely to remember you when you ask them for advice or work experience. For example, multiple career-seminar speakers have sent me details afterwards about their job-search strategy, companies they'd considered and more.

Arrange an internship. Use your (or your supervisor's) network, or cold-e-mail people and ask them for internship or work-shadowing opportunities. In February, I spent one week at *Nature Immunology* (after meeting an editor at a conference), and this was enough to get a good understanding of day-to-day life in the job. I was involved in discussing article submissions, wrote a commentary and identified appropriate reviewers for a paper. In addition, I spoke to people in the company and learnt about their roles: for example, I found out the differences between primary-journal and review-journal editors.

WHAT SUPERVISORS CAN DO

A free and open exploration of different career options requires a work environment that encourages students to take out time for career exploration and that actively helps them to find the relevant information. However, most supervisors arguably know little about non-academic careers and might not have much incentive to provide students with access to such information. As a consequence, only one-third of the respondents in *Nature*'s 2017 Graduate Student Survey reported receiving useful advice from their supervisor regarding non-academic careers (see https://go.nature.com/2qwsyfx).

"Researchers who leave academia are not failed academics."

When I was involved in organizing the 2016 Medical Sciences Careers Day for PhD students at the University of Oxford, some suggested also inviting academic speakers to provide a more 'balanced picture'. But PhD students already have plenty of role models in their academic environment and solid access to information about the academic career track. And the main purpose of events such as this is to provide information that students don't have access to, and thereby correct the already existing imbalance.

PhD students alone are unlikely to transform the culture of academic science. I think that PhD programmes and individual supervisors have a responsibility to make it easy for students to find the path that is right for them, and I know many group leaders who see this as part of their role as a mentor. As a principal investigator (PI), you can take several steps to support your students.

Support students' professional development. Offer students opportunities to teach, supervise, chair meetings, manage collaborations and write papers or grant applications. In my experience, this is already widely practised, but some PIs do not always appreciate the importance of these training opportunities to students. I learnt the most when I had to develop a certain skill — say, doing a practice run for a major presentation — and received feedback. You could formalize this into a regular meeting and discuss what the students enjoy and where they need more practice.

Establish a flexible work environment. The students I have met who had developed a clear career plan by the end of their PhD were usually from research groups that cultivated a flexible, low-pressure environment. Encourage students to take time out for extracurricular activities that will provide them with useful experiences. I have been fortunate to have had the opportunity to try out different things, but I know others who feel pressured to be in the lab all day, every day.

Encourage work experience. Some formal PhD programmes include an option, even a requirement, to do an internship outside academia, and I think this is a great idea. Friends of mine have spent three months drafting policy at the World Health Organization or filming science documentaries for the BBC. Everyone I know who has done an internship greatly valued the experience, and individual supervisors can support this — and let their students know that they do — by, for example, providing contacts in relevant sectors.

Show your support. Arguably the most important thing that supervisors can do is openly show support for lab members and mentees who might be heading out of academia. For example, you could say on your lab page that it is important to you to support your lab members' professional development. You could supplement the statement with an up-to-date list of lab alumni to show that you are proud of their achievements. Potential PhD and postdoc applicants will greatly appreciate your stand on this issue.

A PHD IS HIGHLY VALUABLE

This leaves us with one last aspect of the culture of failure and its effect on doctoral students and postdocs: the widespread misconception that a PhD is useful training only for academic research. Or, in other words, if you leave academia, your mum will think that you've wasted your time doing a PhD. You might even have wondered about that yourself.

We know that most PhD graduates eventually go on to other careers, but have they all wasted their time? Absolutely not. The skills you are acquiring (or have acquired) during a PhD are highly sought by employers beyond academic science. You are incredibly resilient, hard-working and motivated. You make decisions based on evidence, you can interpret data, you can communicate complex concepts clearly, you are an effective team player and you can prioritize tasks. And you have a degree to prove all of this.

You have every reason to be positive about your job prospects.

Personally, I won't regret having done my PhD, regardless of my future career. ■

Philipp Kruger *is a PhD student in immunology at the University of Oxford, UK.*

© 2018 Springer Nature Limited. All rights reserved.

THE TAIL OF DANNY WHISKERS

It's not easy being smart.

BY FAWAZ AL-MATROUK

Ten miles from the Canadian border, they were stopped by police.

"Don't say anything," said Dr Tarboush, his hands turning red against the steering wheel.

"You were driving like a maniac," said the cat in the carrier box on the passenger seat. His name was Danny Whiskers.

Dr Tarboush could feel his heartbeat in his hands as he watched the police officer approach in the rear-view mirror.

"You think I'm stupid?" said Danny. "I'm not stupid."

"Shush!"

"I know when to be quiet."

"Hello officer," said Dr Tarboush.

The police officer kept his distance. "Licence and registration please."

Dr Tarboush fumbled through the glove compartment. It overflowed. Receipts, envelopes, pages from a scientific paper on the epigenetics of intelligence.

"Hang on," said the doctor, searching through papers. "I have it somewhere."

"Goodness," whispered Danny.

"Shush!"

"You'll get us shot."

"You're the talking cat!"

The officer heard him. "Excuse me?"

"Sorry, officer, just talking to myself. Found it!" Dr Tarboush produced the registration.

The officer twisted his brow. "Hang tight."

Dr Tarboush watched the officer get smaller in the rear-view mirror. Beads of sweat collected on his bushy eyebrows. "He's onto us."

"Great. Ten miles from Canada and you'll have me killed."

"You're the one who can't keep quiet."

"You're the one swerving across lanes."

"That's because my cat decided to break out of his carrier box and shout obscenities," — he was loud now — "at my fellow drivers!"

"Do they not understand what a fast lane is?"

"Do you not understand what a fugitive is?"

"I understand. I'm not stupid."

"I get ten years for unauthorized experiments in genetic engineering."

"I know."

"You get to learn how many ways there are to skin one of you."

"Stop."

"I should have left you in the lab and read about you in the *Vancouver Times*, or whatever they have up there, sipping chamomile tea. 'Talking cat!' 'Freak experiment!' 'Puss gives heartfelt monologue as federal science police put him to …' Are you crying?"



ILLUSTRATION BY JACEY

Danny was distinctly crying. "No."

"Stop crying."

"It's not fair," he mumbled. "I didn't ask for this. I didn't ask to be smart. I was legal just a week ago. Why did they have to change the laws? Why can't they just let me be?"

Dr Tarboush squirmed in the driver's seat. He was never good at comforting.

"I don't know, Danny."

"It's not fair."

"No, it's not … People get scared and make laws and they ruin other people's lives and they don't understand how."

Danny sniffled. "Other people's lives?"

"Yeah."

"You mean I'm other people?"

"Of course you are."

Danny let himself smile. "Smarter than human people, though."

"Shove it."

Danny laughed. Dr Tarboush stuck his fingers through the metal grating of the carrier box. "Come here."

Danny leant into the fingers, accepting the neck rubs. He began to purr.

"That's a good boy," said Dr Tarboush.

"Don't patronize me," said Danny, purring.

Heel clicks announced the officer's return. Dr Tarboush sat up straight. The officer leant into the car, a yellow ticket between his fingers. He flung it onto the pile of papers.

"Figured that's where you want it," he said.

"Sure."

"It's for changing lanes without a signal."

Dr Tarboush nodded. "Right."

"That your cat?"

"Yeah."

"Cute."

Danny groaned in his carrier box. Dr Tarboush felt his heart sink into his belly.

The officer smiled. "You know, there's people doing weird experiments on animals. Changing their genes and such. Something called crispy."

Danny scoffed.

Dr Tarboush scoffed to cover it up.

"You scoffing at me?"

"No, no, officer. At crispy," his voice broke. "It's a stupid idea."

The officer narrowed his eyes. "Why stupid?"

Dr Tarboush cleared his throat. "It's against God's plan, you know. There ought to be a law."

"There is a law."

"Is there?"

"Serious law. Serious time. There's two fugitives from it now. An old man and his cat, last seen heading north from California. Talkative cat. Described as 'uppity.'"

"Uppity!" It was Danny Whiskers.

Dr Tarboush slammed his forehead against the steering wheel.

The officer said: "That was the cat."

"Sir …"

"Sounded like the cat."

"I'm a ventriloquist, sir. I can throw voices."

"You're a ventriloquist," the officer laughed. "Dr Taha Tarboush. PhD in ventriloquy. Or is it ventriloquation? I don't know. I don't care. Back to the subject. You say it's against God's plan to mess with genetics. I say," he leant in to whisper, "that my dad has Alzheimer's, and I'm hoping for a breakthrough. Godspeed."

With that, the officer walked away. Dr Tarboush could feel his heartbeats, six to every footstep. "But don't speed!" The officer called from his cruiser.

"What just happened?" asked Danny Whiskers.

"I think we got lucky."

"I think truth won out."

"Yeah."

"I'm truth," said Danny. "I won out."

Dr Tarboush turned to his passenger seat. "We have ten miles to the border, Danny. One more sound, and you're a stray."

"What about a purr?"

"Danny."

"Meow?"

"Shut up."

"What if my tail falls asleep? Can I let you know? Can I leave my box? What if it itches? What if I need a scratch? Scratching's a sound. Can I make a scratch?"

They drove ten miles to the border without a moment of silence between them. Shortly after they entered British Columbia, Danny Whiskers fell asleep. ■

Fawaz Al-Matrouk *loves the latest tech, but writes with pen and paper. He is a film director by trade, currently developing a feature debut with the support of SFFILM.*

© 2018 Springer Nature Limited. All rights reserved.